Postprocessing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests

ERIC D. LOKEN

Cooperative Institute for Mesoscale Meteorological Studies, and School of Meteorology, University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

ADAM J. CLARK

School of Meteorology, University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

AMY MCGOVERN

School of Meteorology, University of Oklahoma, Norman, Oklahoma

MONTGOMERY FLORA

Cooperative Institute for Mesoscale Meteorological Studies, and School of Meteorology, University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

KENT KNOPFMEIER

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 30 May 2019, in final form 19 September 2019)

ABSTRACT

Most ensembles suffer from underdispersion and systematic biases. One way to correct for these shortcomings is via machine learning (ML), which is advantageous due to its ability to identify and correct nonlinear biases. This study uses a single random forest (RF) to calibrate next-day (i.e., 12–36-h lead time) probabilistic precipitation forecasts over the contiguous United States (CONUS) from the Short-Range Ensemble Forecast System (SREF) with 16-km grid spacing and the High-Resolution Ensemble Forecast version 2 (HREFv2) with 3-km grid spacing. Random forest forecast probabilities (RFFPs) from each ensemble are compared against raw ensemble probabilities over 496 days from April 2017 to November 2018 using 16-fold cross validation. RFFPs are also compared against spatially smoothed ensemble probabilities since the raw SREF and HREFv2 probabilities are overconfident and undersample the true forecast probability density function. Probabilistic precipitation forecasts are evaluated at four precipitation thresholds ranging from 0.1 to 3 in. In general, RFFPs are found to have better forecast reliability and resolution, fewer spatial biases, and significantly greater Brier skill scores and areas under the relative operating characteristic curve compared to corresponding raw and spatially smoothed ensemble probabilities. The RFFPs perform best at the lower thresholds, which have a greater observed climatological frequency. Additionally, the RF-based postprocessing technique benefits the SREF more than the HREFv2, likely because the raw SREF forecasts contain more systematic biases than those from the raw HREFv2. It is concluded that the RFFPs provide a convenient, skillful summary of calibrated ensemble output and are computationally feasible to implement in real time. Advantages and disadvantages of ML-based postprocessing techniques are discussed.

1. Introduction

Over the past 20 years, increases in computing resources have reshaped the state of numerical weather

Corresponding author: Eric D. Loken, eric.d.loken@noaa.gov

DOI: 10.1175/WAF-D-19-0109.1

prediction (NWP) in several key ways: by enabling skillful high-resolution ensemble forecasts (e.g., Xue et al. 2007; Jirak et al. 2012, 2016, 2018; Roberts et al. 2019; Clark et al. 2018; Schwartz et al. 2015, 2019); by increasing the capacity to run and store models for research and operations (e.g., Hamill and Whitaker 2006; Kain et al. 2010; Hamill et al. 2013; Clark et al. 2018; Roberts et al. 2019); and by reducing the time required to perform complex analyses, enabling more—and more frequent—high-resolution NWP products (e.g., Kain et al. 2010; Gallo et al. 2017, 2019; Roberts et al. 2019). These changes have led to large improvements in NWP quality and value, particularly for fields related to convection. For example, the higher resolution associated with convection-allowing models (CAMs; i.e., those that explicitly simulate convection and run with horizontal grid spacing $\leq \sim 4$ km) has improved forecasts of storm initiation, evolution, and mode compared to convection-parameterizing models (e.g., Kain et al. 2006). Meanwhile, convection-allowing ensembles (CAEs) provide further benefits by accounting for uncertainties in initial conditions and/or model physics (e.g., Roebber et al. 2004; Leutbecher and Palmer 2008; Clark et al. 2009) and conveying forecast uncertainty information to the user (e.g., Palmer 2017). Despite ensembles' higher computational cost, their benefits have been well documented at both convection-parameterizing (e.g., Epstein 1969; Leith 1974; Du et al. 1997; Stensrud et al. 1999; Wandishin et al. 2001; Bright and Mullen 2002; Clark et al. 2009) and convection-allowing (e.g., Coniglio et al. 2010; Loken et al. 2017; Schwartz et al. 2017) resolutions.

Nevertheless, CAMs and CAEs still have biases in the placement, timing, and magnitude of precipitation-producing weather systems (e.g., Davis et al. 2006; Kain et al. 2008; Weisman et al. 2008; Herman and Schumacher 2016, 2018c). Additionally, CAEs remain relatively expensive to run and thus typically have small ensemble membership (e.g., Schwartz et al. 2014; Clark et al. 2018). While small ensembles (e.g., consisting of 10–30 members) have been found to deliver nearly as much forecast skill as larger ensembles (e.g., up to 50 members; Clark et al. 2011; Schwartz et al. 2014; Sobash et al. 2016), they can undersample the forecast probability density function (PDF; e.g., Schwartz et al. 2010, 2014; Roberts et al. 2019), potentially leading to degraded reliability and underdispersion, especially in the absence of neighborhood evaluation or post-processing methods (Schwartz et al. 2014). Indeed, most CAMs and CAEs are currently underdispersive (e.g., Romine et al. 2014). One method to increase CAE spread is to increase the diversity of the ensemble membership, which can be achieved by using members with multiple dynamic cores, analyses, boundary layer schemes, microphysics parameterizations, and initialization periods [e.g., the Storm-Scale Ensemble of Opportunity (Jirak et al. 2012, 2016) and the High-Resolution Ensemble Forecast System, version 2 (Jirak et al. 2018; Roberts et al. 2019)]. While diverse, informally designed ensembles can produce skillful forecasts (Jirak et al. 2016, 2018; Clark et al. 2018; Schwartz et al. 2019), their skill comes with several notable drawbacks. One is that the ensemble members tend to cluster around multiple solutions based on their dynamic core (e.g., Schwartz et al. 2019). This member clustering can cause the ensemble mean forecast to fall outside of the clusters of member solutions (see Fig. 1 in Schwartz et al. 2019) and can adversely affect the quality of the ensemble probabilities, since each member's solution is not equally likely to occur (Schwartz et al. 2019). Another potential consequence of multimodel, multiple-physics CAEs is an artificial inflation of ensemble spread due to the existence of systematic biases between ensemble members (Eckel and Mass 2005; Clark et al. 2010b; Loken et al. 2019). These shortcomings are typically resolved using one or more postprocessing techniques, including isotropic (e.g., Sobash et al. 2011, 2016; Loken et al. 2017, 2019; Roberts et al. 2019) or anisotropic (e.g., Marsh et al. 2012) spatial smoothing of the raw forecast probability field, recalibration of forecast probabilities (e.g., Hamill et al. 2008), probability matching techniques (e.g., Ebert 2001; Clark et al. 2010a,b; Clark 2017; Loken et al. 2019), and various neighborhood-based methods to construct ensemble probabilities (e.g., Schwartz et al. 2010; Blake et al. 2018; Roberts et al. 2019; Schwartz and Sobash 2017).

Another avenue for postprocessing is machine learning (ML; e.g., McGovern et al. 2017). Conceptually, ML algorithms identify patterns in historical data and use these patterns to correct for systematic ensemble biases. This idea is not new; dynamical–statistical methods have existed since at least the 1950s (e.g., Malone 1955; Klein et al. 1959). One example of a well-performing traditional technique is Model Output Statistics (MOS; Glahn and Lowry 1972), which relates NWP output to observed variables of interest (e.g., observed precipitation). ML-based postprocessing methods work similarly; however, while MOS techniques tend to be based on linear regression (e.g., Glahn and Lowry 1972), ML techniques are not necessarily linear. A variety of ML approaches, other than regression, have been applied to weather prediction since the 1980s and include: artificial neural networks (ANNs; e.g., Key et al. 1989; Marzban and Stumpf 1996; Kuligowski and Barros 1998; Hall et al. 1999; Manzato 2007; Rajendra et al. 2019), support vector machines (e.g., Ortiz-García et al. 2014; Adrianto et al. 2009),

clustering algorithms (e.g., Baldwin et al. 2005), genetic algorithms (e.g., Szpiro 1997; Kishtawal et al. 2003; Wong et al. 2008), and decision tree–based methods (Breiman 1984, 2001; Herman and Schumacher 2018c).

Although the ML algorithms mentioned above are not ''new''—the random forest (RF) technique utilized herein was described nearly 20 years ago by Breiman (2001)—enhanced computing power and storage capacity have facilitated the successful application of ML to NWP in recent years (e.g., McGovern et al. 2017, and works cited therein). Indeed, as computing power and storage continue to increase, the role ML plays in NWP postprocessing is likely to grow as well. Especially as forecasters confront an ever-increasing deluge of data (e.g., Carley et al. 2011; McGovern et al. 2017; Karstens et al. 2018), ML or other postprocessing techniques may be desired to quickly and effectively summarize information from NWP products. Therefore, this paper seeks to address important basic questions regarding the application of ML techniques in general—and the RF algorithm in particular—to NWP postprocessing. Considerations include what, if anything, a ML approach provides relative to simpler forms of postprocessing (e.g., two-dimensional spatial smoothing) and how feasible it would be to implement ML-based predictions operationally. Specifically, the costs and benefits of an RF-based approach are considered relative to two-dimensional isotropic spatial smoothing for two multimodel, multi-analyses, multiphysics ensembles: the convection-parameterizing Short-Range Ensemble Forecast System (SREF; Du et al. 2015) and the convection-allowing High-Resolution Ensemble Forecast System, version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019). A focus on precipitation is adopted herein due to its importance as a sensible weather field related to convection and the high economic and human impacts of heavy precipitation events (e.g., NCEI 2019). The next-day (i.e., 1200 UTC–1200 UTC) time frame is selected due to its relative simplicity and to match operational Day 1 products issued by the Weather Prediction Center (WPC).

The remainder of this paper is organized as follows: section 2 details the methods and datasets used herein, section 3 describes the results and presents two case studies for analysis, section 4 summarizes and discusses important findings, and section 5 concludes the paper and outlines avenues for future work.

2. Methods

a. Datasets

Forecast data from the SREF and HREFv2 are considered over 496 common days, spanning April 2017–November 2018 (Table 1). The analysis domain for both ensembles covers the contiguous United States (CONUS; Fig. 1), and the analysis period covers 24 h (1200 UTC–1200 UTC the next day). Details on each ensemble's configuration are given below.

The SREF is a 26-member convection-parameterizing ensemble in which half of the members use the Advanced Research Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) dynamic core and half use the dynamic core from the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). The SREF uses 16-km horizontal grid spacing and runs four cycles per day at 0300, 0900, 1500, and 2100 UTC (Du et al. 2015), with forecast fields output every 3 h. This study uses 15–39-h forecasts from the 2100 UTC initialization. Due to storage and data availability constraints, the SREF analyses herein are output to a grid with 32-km horizontal grid spacing (NCEP grid 221). SREF configuration details are summarized in Table 2.

The HREFv2 originates from the Storm Prediction Center's Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016, 2018), which was developed as a collection of individual CAMs with different dynamic cores, analyses, initialization times, microphysics, and boundary layer parameterizations. Although the HREFv2 and SSEO use ad hoc, informal designs, they have consistently outperformed other CAEs (Jirak et al. 2016, 2018; Schwartz et al. 2019). Indeed, the strong performance of the HREFv2 led to its implementation as the National Weather Service's first operational CAE on 1 November 2017 (Jirak et al. 2018; Roberts et al. 2019). Despite the drawbacks arising from its informal design (e.g., unequal likelihood, member clustering, maintenance difficulties; Schwartz et al. 2019), it remains a ''high-quality baseline'' (Schwartz et al. 2019) for CAE performance.

The HREFv2 comprises eight members, with half the membership composed of 12-h time lagged runs (Jirak et al. 2018; Roberts et al. 2019). The nonlagged (time-lagged) members are initialized daily at 0000 UTC (the previous day at 1200 UTC). All members use approximately 3-km horizontal grid spacing and collectively contain two dynamic cores, two microphysics schemes, and two boundary layer parameterizations. Forecast fields are output hourly from each member. 12–36-h HREFv2 forecasts are used herein. Full details of HREFv2 configuration are given in Table 3.

National Center for Atmospheric Research/Earth Observing Laboratory (NCAR/EOL) Stage IV precipitation data (Lin 2011) are used for observations. While the dataset has known deficiencies, especially in regions of complex terrain where radar coverage is sparse and/or inaccurate (e.g., Hitchens et al. 2013; Herman and

TABLE 1. Forecast valid dates for each ensemble.

Month	2017	2018
January	—	1–31
February	—	1–28
March	—	1–10, 14–17, 19–20, 22–26
April	28	7–30
May	1–2, 4–5, 7–10, 13–23, 26–31	1–31
June	1, 6–7, 9, 11–13, 15, 17–25	1–7, 10–30
July	3–6, 15–16, 18–19, 22–24, 30–31	1–31
August	1–10, 12–15, 17–30	1–5, 8–31
September	1–10, 13–15, 17–30	1–30
October	1–24, 26–31	1–31
November	1–30	1–4, 6, 9–13
December	1–31	—

Schumacher 2016, 2018b), the dataset has high-resolution (~4.8-km grid spacing) coverage over the full CONUS, making it the preferred observational dataset.

b. *Obtaining raw and spatially smoothed ensemble forecasts*

Raw SREF and HREFv2 forecast probabilities are computed by first remapping each member's 24-h (1200 UTC–1200 UTC) quantitative precipitation forecast to NCEP grid 215, which has approximately 20-km horizontal grid spacing. The remapping is done using a neighbor budget method (Accadia et al. 2003), a nearest-neighbor averaging method that approximately conserves total precipitation. Upscaling to 20 km saves significant computational expense and better matches scales at which predictability should exist at 12–36-h lead times. After upscaling, the fraction of ensemble members exceeding a given precipitation threshold is calculated at each point on the 20-km grid. Four 24-h precipitation thresholds are considered: 0.1, 0.5, 1, and 3 in. (i.e., 2.54, 12.7, 25.4, and 76.2 mm).

Given the underdispersive properties of most CAEs, a two-dimensional, isotropic Gaussian kernel density function (e.g., Sobash et al. 2011, 2016; Loken et al. 2017, 2019; Roberts et al. 2019) is often applied to a CAE's raw forecast probability field as a simple but effective means of increasing forecast spread and reducing overforecasting bias. Since most CAEs are overconfident and underdispersive, spatial smoothing typically enhances reliability and resolution, but oversmoothing can degrade reliability and sharpness (Sobash et al. 2011, 2016; Loken et al. 2017, 2019; Roberts et al. 2019). In this study, as in Loken et al. (2019), the following equation is applied to the (remapped) SREF and HREFv2 raw ensemble forecast probabilities to create isotropic spatially smoothed forecast probabilities:

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2\right], \quad (1)$$

where f is the forecast probability at a given point, N is the number of points where at least one ensemble

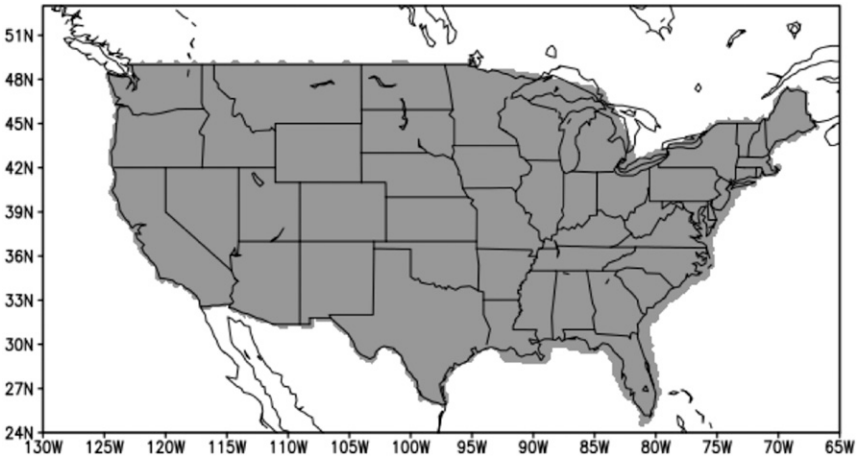

FIG. 1. Analysis domain for each ensemble.

TABLE 2. SREF member specifications, adapted from Du et al. (2015). Initial conditions (ICs) are taken from the operational Rapid Refresh (RAP; Benjamin et al. 2016), the National Centers for Environmental Prediction's (NCEP's) Global Forecast System (GFS), and the North American Mesoscale Model Data Assimilation System (NDAS). IC perturbations are derived using a blend of Global Ensemble Forecast System (GEFS) and SREF analyses. Lateral boundary conditions (LBCs) are from the GFS and GEFS members. Convective parameterizations include the Kain–Fritsch (KF; Kain 2004), Grell (1993), Betts–Miller–Janjić (BMJ; Betts 1986; Janjić 1994), and simplified Arakawa–Schubert (Han and Pan 2011) schemes. Planetary boundary layer (PBL) schemes include the Yonsei University (YSU; Hong et al. 2006), Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004, 2006), Mellor–Yamada–Janjić (MYJ; Janjić 2002) parameterizations as well as that used in the GFS. Microphysics schemes include the WRF single-moment 6-class (WSM6; Hong and Lim 2006), Thompson et al. (2004), and Ferrier et al. (2002) schemes.

Member	ICs	LBCs	Convective	PBL	Microphysics
arw_ctl	RAP	GFS	KF	YSU	WSM6
arw_p1	RAP	GEFS13	Grell	MYNN	Thompson
arw_n1	RAP	GEFS14	BMJ	MYJ	Ferrier
arw_p2	RAP	GEFS15	BMJ	MYJ	Thompson
arw_n2	RAP	GEFS16	KF	YSU	Ferrier
arw_p3	GFS	GEFS17	KF	YSU	Thompson
arw_n3	GFS	GEFS18	Grell	MYNN	WSM6
arw_p4	GFS	GEFS19	KF	YSU	Ferrier
arw_n4	GFS	GEFS20	BMJ	MYJ	WSM6
arw_p5	NDAS	GEFS1	KF	YSU	WSM6
arw_n5	NDAS	GEFS2	Grell	MYNN	Ferrier
arw_p6	NDAS	GEFS3	Grell	MYNN	Thompson
arw_n6	NDAS	GEFS4	BMJ	MYJ	Thompson
nmmb_ctl	NDAS	GFS	BMJ	MYJ	Ferrier hi-res
nmmb_p1	NDAS	GEFS1	BMJ	MYJ	Ferrier hi-res
nmmb_n1	NDAS	GEFS2	SAS	GFS	WSM6
nmmb_p2	NDAS	GEFS3	BMJ	MYJ	WSM6
nmmb_n2	NDAS	GEFS4	SAS	GFS	Ferrier hi-res
nmmb_p3	GFS	GEFS5	BMJ	MYJ	WSM6
nmmb_n3	GFS	GEFS6	SAS	GFS	Ferrier hi-res
nmmb_p4	GFS	GEFS7	BMJ	MYJ	Ferrier hi-res
nmmb_n4	GFS	GEFS8	SAS	GFS	WSM6
nmmb_p5	RAP	GEFS9	BMJ	MYJ	Ferrier hi-res
nmmb_n5	RAP	GEFS10	SAS	GFS	WSM6
nmmb_p6	RAP	GEFS11	BMJ	MYJ	WSM6
nmmb_n6	RAP	GEFS12	SAS	GFS	Ferrier hi-res

member exceeds the given precipitation threshold, d_n is the distance from the current point to the nth point, and σ is the standard deviation of the Gaussian kernel. Importantly, σ controls the degree of spatial smoothing and must be tuned appropriately to produce skillful forecasts. Herein, σ is chosen such that the resulting collection of daily, CONUS-wide forecast probabilities minimizes the Brier score (BS;

TABLE 3. HREFv2 member specifications. HRW and NAM refer to High Resolution Window and North American Mesoscale Model runs, respectively. The "−12 h" in the first column indicates a 12-h time lagged member (i.e., 1200 UTC initialization the previous day instead of 0000 UTC initialization). Initial conditions and lateral boundary conditions (IC/LBCs) are taken from the NAM, Rapid Refresh (RAP), and/or Global Forecast System (GFS), as indicated. A "−6 h" indicates that the model from which the IC/LBCs are derived was initialized 6-h before the given HREFv2 member. Microphysics schemes include the WRF single-moment 6-class (WSM6; Hong and Lim 2006) and the Ferrier–Aligo (Aligo et al. 2018) schemes, while boundary layer parameterizations include the Mellor–Yamada–Janjić (MYJ; Janjić 2002) and Yonsei University (YSU; Hong et al. 2006) schemes.

Member	Model Core	IC/LBCs	Microphysics	PBL
HRW NSSL	WRF-ARW	NAM/NAM −6 h	WSM6	MYJ
HRW NSSL −12 h	WRF-ARW	NAM/NAM −6 h	WSM6	MYJ
HRW ARW	WRF-ARW	RAP/GFS −6 h	WSM6	YSU
HRW ARW −12 h	WRF-ARW	RAP/GFS −6 h	WSM6	YSU
HRW NMMB	NMMB	RAP/GFS −6 h	Ferrier–Aligo	MYJ
HRW NMMB −12 h	NMMB	RAP/GFS −6 h	Ferrier–Aligo	MYJ
NAM CONUS Nest	NMMB	NAM/NAM	Ferrier–Aligo	MYJ
NAM CONUS Nest −12 h	NMMB	NAM/NAM	Ferrier–Aligo	MYJ

e.g., Wilks 1995) over the training dataset. The BS can be expressed as

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{2}$$

where N is the total number of forecast–observation pairs (i.e., the number of grid points in the domain multiplied by the number of days in the dataset), f_i is the forecast probability at the ith grid point, and o_i is the binary observation at the ith grid point.

c. Random forest–based forecasts

While the umbrella of machine learning includes many popular and powerful algorithms, the random forest (Breiman 2001) algorithm has some important advantages that make it the preferred technique in this study. Namely, RFs do not require standardized inputs, they have relatively few hyperparameters to tune, they are parallelizable and thus relatively fast to run, and previous studies (e.g., Gagne et al. 2014; Herman and Schumacher 2018a,c) have found that they perform well for precipitation prediction.

The building blocks of RFs are individual decision trees (Breiman 1984). Decision trees recursively split a dataset by selecting, at each node, the variable and threshold that maximizes a dissimilarity metric (e.g., information gain) until a stopping criterion is reached (e.g., the number of dataset samples falls below a specified amount, the tree reaches a certain depth, etc.). Once the splitting criteria are determined for each node using the training data, the tree can be used for prediction on a testing dataset by sorting testing samples through the tree. Testing probabilities are given by the fraction of training samples associated with an observed event of interest at the terminal node, or "leaf node," into which a testing sample is classified. One drawback of individual decision trees is that they tend to be overly sensitive to small variations in the training dataset (e.g., Gagne et al. 2014). RFs provide a solution to this so-called "brittleness" (Gagne et al. 2014) by growing multiple trees, which are unique due to the introduction of stochasticity into the training process. Specifically, each tree in the RF uses a subset of training samples determined by bootstrap resampling (i.e., resampling with replacement; e.g., Wilks 1995) the full set, and splits at each node are determined by considering a random subset of variables. In the RF framework, testing probabilities of event occurrence are simply the mean testing probabilities from each tree. Although the RF's multiple trees may make it more difficult for humans to interpret RF output probabilities, the RF method is generally attractive since it is resistant to overfitting and tends to produce outputs with low bias (e.g., Breiman 2001). More details on the RF technique can be found in Herman and Schumacher (2018c), McGovern et al. (2017), and Gagne et al. (2014).

Herein, 18 (20) fields are used as inputs into the RF algorithm to obtain SREF (HREFv2) RFFPs (Table 4). These fields include variables that represent a point's meteorological environment, variables that have an obvious direct relationship with observed precipitation, and latitude and longitude, which are designed to account for spatially varying precipitation climatology. Simulated 2–5-km updraft helicity (UH) is also included as a predictor given its relationship to sustained rotating updrafts and severe weather occurrence (e.g., Kain et al. 2008; Sobash et al. 2011; Loken et al. 2017), since supercells or mesoscale convective systems that produce elevated values of simulated UH may also produce localized heavy rainfall (e.g., Nielsen and Schumacher 2018). The SREF uses two less fields compared to the HREFv2 since the SREF does not output forecasts of simulated reflectivity or UH.

Predictors are derived from ensemble forecast gridpoint values on the 20-km grid. Originally, predictors included forecasts from each ensemble member, since it was hypothesized that the RF algorithm could learn and correct for each member's individual systematic biases. However, simply using the ensemble mean value of each variable produced RFFPs that were at least as skillful as those made using predictors from each member. Moreover, using only ensemble mean forecast values made it computationally feasible for the RF to consider predictors from multiple points in space, potentially allowing the RF to identify and correct nonlinear systematic spatial biases. Therefore, ensemble mean forecast values from points (on the 20-km grid) within an approximately 100-km box surrounding the forecast point (i.e., forecast values from the forecast point and the 24 closest points) are used as predictors. Notably, there is no spatial averaging of the values used beyond the neighbor budget interpolation to the 20-km grid.

Further necessary reductions in dataset dimensionality are achieved through preprocessing the raw ensemble data. First, a temporal mean is taken over the 8 three-hourly (24 one-hourly) forecast fields each day at each native grid point for the SREF (HREFv2). While useful information is undoubtedly lost using this method, the temporal mean provides an overall summary of the simulated meteorological conditions during the relevant 24-h period, which is hypothesized to be sufficient for skillful RF probabilistic precipitation forecasts on next-day time scales. Each day's temporal mean forecasts are then remapped to the 20-km

TABLE 4. Predictor variables from each ensemble. Asterisks denote variables used for the HREFv2 RFFPs only. Due to limited computing resources, all predictors except for latitude and longitude represent 24-h temporal mean ensemble mean quantities.

Predictor Variable	Atmospheric Level
Temperature	500, 700, and 850 hPa, and 2 m AGL
Dewpoint temperature	500, 700, and 850 hPa, and 2 m AGL
Max hourly simulated reflectivity*	1 km AGL
CAPE	Surface based
CIN	Surface based
Precipitable water (PWAT)	Entire column
Max hourly simulated UH*	2–5 km AGL
Max hourly *U*, *V* wind	10 m AGL
Max hourly upward vertical velocity (UVV), downward vertical velocity (DVV)	100–1000 hPa (400–1000 hPa for NAM members of HREFv2)
Forecast 24-h precipitation	Surface
Lat, lon	—

verification grid. Finally, 10% (i.e., 2130) of the (remapped) points in the analysis domain are randomly sampled without replacement and added to the dataset for training each day (note that the full domain is still used for testing).

Randomly sampling the domain in this manner, as in Gagne et al. (2014), accomplishes two main objectives: it reduces the computational expense of the algorithm by appreciably shrinking the size of the training dataset, and it decreases the likelihood of including multiple highly correlated grid points in the training set, reducing the chance of RF overfitting (i.e., fitting on noise rather than actual, systematic patterns in the data). A sampling rate of 10% is greater than that used by Gagne et al. (2014) but is chosen to balance the trade-off between computational expense and RFFP skill, which increased only slightly at sampling rates beyond 10% in sensitivity tests from 0.5% to 70% (not shown). All data preprocessing steps are summarized in Fig. 2.

After the data has undergone preprocessing, a random forest classifier from the Python module Scikit-Learn (Pedregosa et al. 2011) is used to train the ensemble RFs and create RFFPs. Based on hyperparameter sensitivity tests (not shown), the random forest classifier requires: 200 trees, a maximum tree depth of 15 levels, at least 20 samples per leaf node, the minimization of entropy for splits, and the consideration of $\sqrt{n}$ predictors (where n is the total number of predictors in the dataset) at each node. Separate RFs are trained for each precipitation threshold, but all RFs use the same hyperparameters. Importantly, since each threshold forecast is created independently, there is no guarantee of consistency between the probabilities of different threshold exceedance. However, the use of different RFs for different thresholds enables a more direct comparison of how the RF technique performs at each threshold individually and allows for different types of precipitation events to be predicted from trees/forests with different, potentially more appropriate structures.

Unlike many previous studies (e.g., Gagne et al. 2014; Herman and Schumacher 2018c), separate RFs are *not* trained for each season and/or geographic region. Using a single RF to represent the entire CONUS year-round likely sacrifices forecast skill, since locations have different time- and space-varying climatologies (e.g., Schumacher and Johnson 2006). However, using a single RF considerably simplifies the prediction and maintenance processes of RF-based postprocessing. For example, with multiple regional RFs, RFFPs may be unphysically discontinuous near the border of two regions, requiring additional postprocessing. Moreover, multiple RFs require more computing power to train (or retrain) and run when making daily predictions. Additionally, it is hypothesized that the inclusion of latitude and longitude coordinates as well as seasonally varying environmental variables (e.g., temperature) may help a single RF implicitly account for time- and space-varying precipitation climatologies. This single-RF approach, while perhaps less efficient than a multi-RF approach with explicit dataset filtering, may be advantageous for precipitation prediction since spatially and seasonally distant training data (e.g., forecast precipitation) may have at least some relevance for all forecast points. However, the single-RF approach may be less appropriate to use in problem domains where distant training data are less relevant to a given forecast point.

d. Verification

Sixteenfold cross validation with 31 days per fold is used to verify the forecasts. Verification metrics are computed on the full set of 496 forecasts derived from each fold's testing set. To facilitate a fair comparison between the RFFPs and spatially smoothed forecasts,

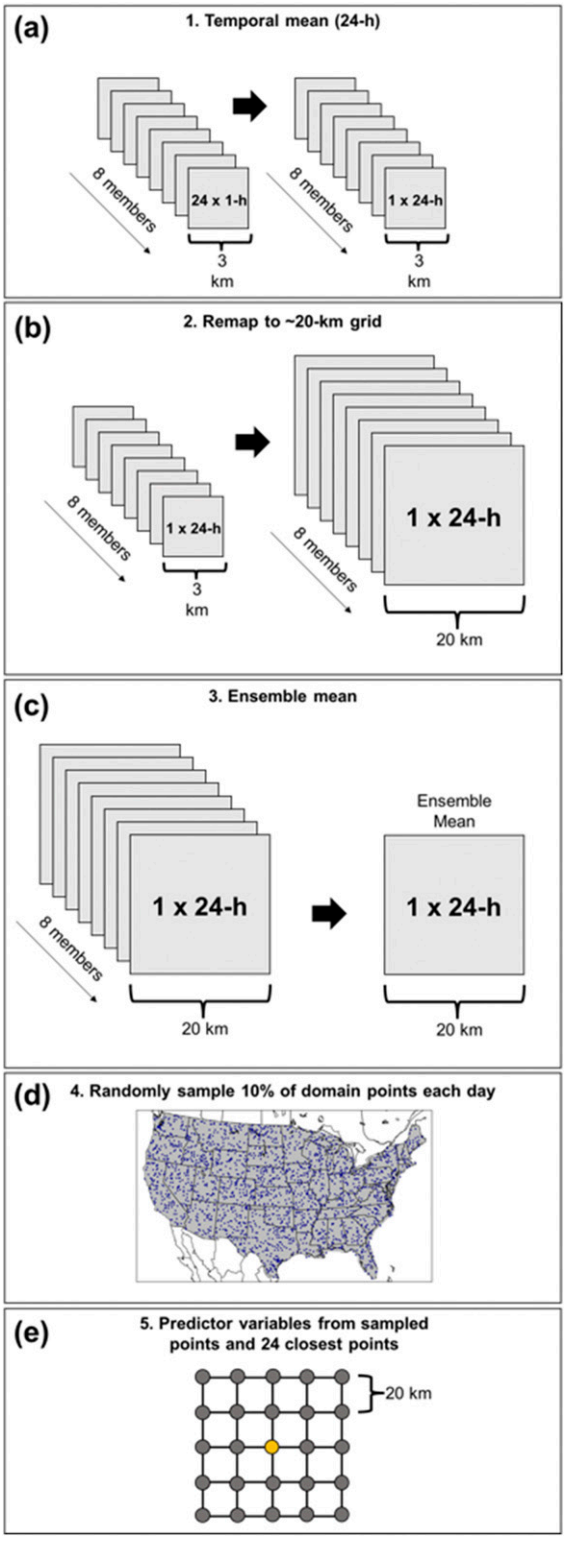

FIG. 2. Schematic illustrating the data preprocessing steps for the 8-member HREFv2. Note that the SREF follows a similar procedure but has 26 members and starts on a coarser native grid. (a) The temporal mean is taken over 24 h at each native grid point for each ensemble member. (b) The temporally averaged data are remapped to an approximately 20-km grid. (c) An ensemble mean is taken at each 20-km grid point. (d) 10% of the domain is randomly sampled for training. (e) Training data consist of the predictor variables at each sampled point (yellow) and the 24 closest 20-km points.

the σ that minimizes the BS over each fold's training set is used to create the spatially smoothed forecasts; hence, σ varies by fold (Fig. 3). Verification metrics are computed over the full domain (Fig. 1) as well as over five distinct regions (Fig. 4), which are based on combinations of the regions defined by Bukovsky (2011). These regions have distinct temperature and precipitation climatologies.

An important strategy for evaluating probabilistic forecasts is the creation of 2×2 contingency tables (e.g., Wilks 1995), which are derived from binarizing the forecast at various probability thresholds. Verification metrics such as probability of detection (POD), probability of false detection (POFD), success ratio (SR), bias, and critical success index (CSI) can then be obtained [e.g., see Eqs. (3)–(7) in Loken et al. 2017]. These metrics form the basis of other forecast evaluation tools used herein, such as the ROC curve (Mason 1982) and performance diagram (Roebber 2009). ROC curves plot POD against POFD at multiple forecast probability thresholds (here, 1%, 2%, and 5%–95% in intervals of 5%). Area under the ROC curve (AUC) provides a measure of forecast discrimination ability, with values of 1 (0.5) indicating a perfect (random) forecast. Since AUC is not sensitive to forecast reliability (Wilks 2001), attributes diagrams (Hsu and Murphy 1986; Wilks 1995) measure reliability by grouping forecasts into k bins based on forecast probability and plot the mean observed relative frequency of each bin against the bin's probability. Herein, 11 bins are used [0%, 5%), [5%–15%), ... , [85%–95%), and [95%–100%]. Perfectly reliable forecasts fall along a diagonal line with a slope of 1 passing through the origin. Over- (under-) forecasts fall below (above) the perfect reliability line. Horizontal and vertical lines are plotted at the sample climatological relative frequency, while a "no skill" line is plotted halfway between the horizontal climatology line and the line of perfect reliability. Points above (below) the no-skill line contribute positively (negatively) to the Brier skill score when a reference forecast of climatology is used (Wilks 1995).

Performance diagrams (Roebber 2009) plot POD against SR and include lines of constant bias and CSI. Herein, forecasts are plotted on performance diagrams at each of the 21 probability levels used to create the ROC curves. The most skillful forecasts fall closest to

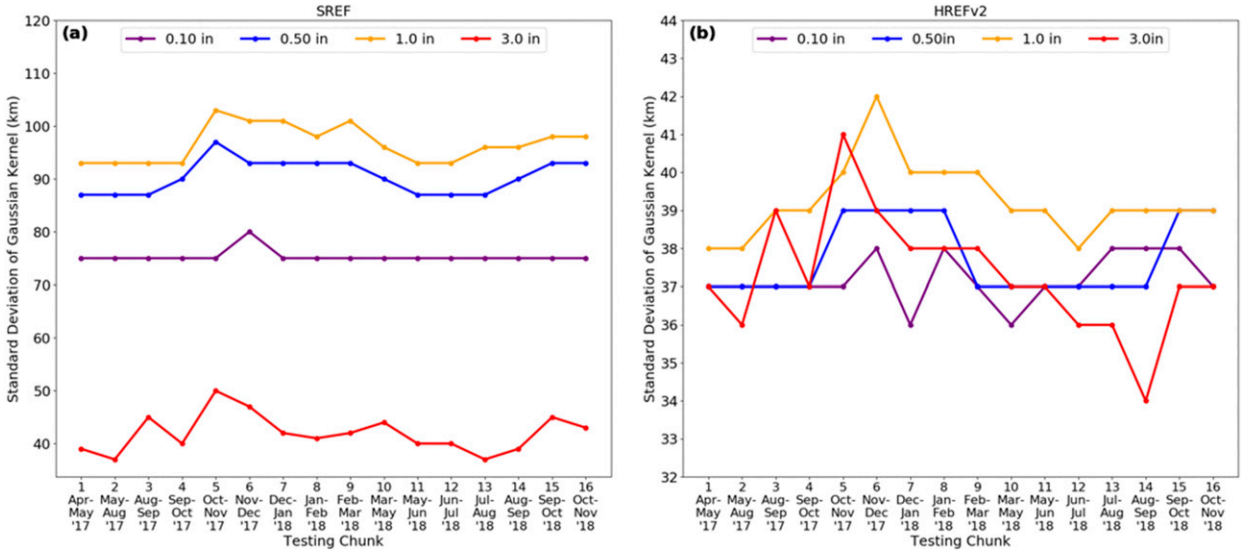

FIG. 3. Relationship between the standard deviation of the Gaussian kernel (i.e., σ) and testing fold for (a) the SREF and (b) the HREFv2. In each plot, 0.1-, 0.5-, 1-, and 3-in. forecasts are depicted in purple, blue, gold, and red, respectively. The range of dates included in each fold is listed on the x axis. Note the different y-axis scales.

the upper right-hand corner of the plot, where POD, SR, bias, and CSI are all optimized.

The BS (e.g., Wilks 1995) measures the magnitude of the forecast probability errors and can be decomposed into reliability, resolution, and uncertainty components (Murphy 1973; Wilks 1995). The BS is a negatively oriented score, so a score of 0 (1) indicates perfect (no) skill. One disadvantage of the BS is that it is sensitive to the observed climatological frequency of the event being verified. The Brier skill score (BSS) helps account for this effect by comparing the BS to that of a reference forecast, which is often a forecast of climatology. The BSS is defined as

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{3}$$

where, herein, $\mathrm{BS}_{\mathrm{ref}}$ is the BS obtained by always forecasting the underlying climatological frequency associated with the entire dataset. The BSS is a positively oriented score, with possible values from $-\infty$ to 1. A BSS of 0 (1) indicates no (perfect) skill relative to the reference forecast.

A one-sided paired permutation test (e.g., Good 2006) is used herein to test whether the AUC and BSS of a given set of forecasts (e.g., the RFFPs) is significantly greater than a second set of forecasts (e.g., the spatially

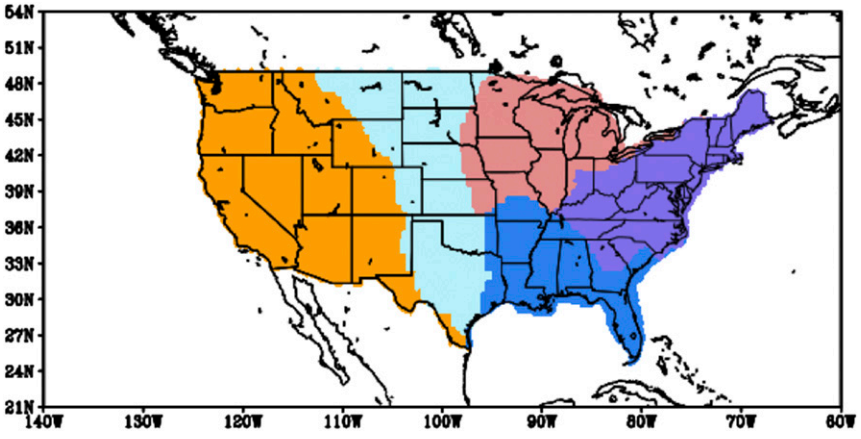

FIG. 4. The five regional analysis regions, which include the West (gold), Great Plains (light blue), Upper Midwest (salmon), South (royal blue), and East (purple).

smoothed probabilities). The general procedure is the same for both AUC and BSS. Individual-day forecasts are randomly permuted between the two forecast systems 10 000 times to create a null distribution of metric differences. The actual difference between the two forecast systems' skill metrics is then compared to the null distribution to obtain a p value. In the AUC paired permutation test, contingency table elements are randomly permuted rather than the AUC values themselves since individual-day AUC values can be very sensitive to small changes in contingency table elements (Hamill 1999). The final AUC values (and AUC differences) for each iteration are computed based on the permuted contingency table elements. In the same manner, individual-day BSs rather than BSSs are permuted, and BSSs (and BSS differences) for each iteration are computed based on the collective permuted BSs.

Spatial biases are assessed using an approach outlined by Clark et al. (2010a) and Marsh et al. (2012). Conceptually, whenever a yes forecast is issued within the domain, the spatial distribution of yes observations within a 500 km × 500 km box is tabulated relative to the yes forecast point and the results are composited over the entire dataset. However, these yes observations are only added to the composite if they fall within the analysis domain. While this method can yield artificially anisotropic contributions to the composite near the domain boundaries, tests (not shown) have indicated that, overall, this method does not appreciably bias the center of the distribution. Thus, in the absence of systematic spatial biases, the center of the distribution should be located at the yes forecast point.

In this study, a yes observation is defined as the Stage IV data exceeding a quantitative precipitation threshold (e.g., 0.1, 0.5, 1, or 3 in.) on the verification grid, while a yes forecast is defined as the forecast exceeding a probability threshold that, to the nearest 1%, optimizes frequency bias. Defining yes forecasts in this way allows for a clean comparison between forecasts by removing bias magnitude but still allowing for spatial biases. Table 5 shows the forecast probability thresholds and their corresponding frequency biases.

One drawback of ML-based postprocessing techniques is that they assume the underlying dynamical models do not change with time and must be retrained whenever developers implement changes. An important question, therefore, is: how long of a dataset is required for ML to perform adequately? To address this question, RFs are retrained and reevaluated using a dataset comprising the first 62, 124, 248, and 372 days (i.e., the first 1/8, 1/4, 1/2, and 3/4) of the full dataset, respectively.

TABLE 5. Forecast probability thresholds used to (approximately) optimize frequency bias for each forecasting system at each precipitation threshold. Actual values of frequency bias are reported in the fourth column.

Precipitation threshold	Ensemble/ forecast	Forecast probability threshold (%)	Frequency bias
0.1 in.	SREF, raw	62	1.029
	HREF, raw	38	1.040
	SREF, smooth	55	1.011
	HREF, smooth	43	0.996
	SREF, RF	44	0.991
	HREF, RF	43	1.007
0.5 in.	SREF, raw	47	0.957
	HREF, raw	38	0.896
	SREF, smooth	38	0.990
	HREF, smooth	35	0.989
	SREF, RF	33	0.992
	HREF, RF	35	0.987
1 in.	SREF, raw	35	1.049
	HREF, raw	26	1.139
	SREF, smooth	29	1.007
	HREF, smooth	29	1.005
	SREF, RF	26	1.011
	HREF, RF	28	0.998
3 in.	SREF, raw	20	1.045
	HREF, raw	26	0.812
	SREF, smooth	18	0.990
	HREF, smooth	20	1.022
	SREF, RF	17	0.970
	HREF, RF	20	1.030

These RFs use the same hyperparameters as described previously. Although this approach is suboptimal, sensitivity tests suggest that the BSS varies only slightly with different hyperparameters; moreover, the set of hyperparameters used previously was deemed close enough to optimal to make using a constant set of hyperparameters worth the reduced computational expense. As with the full dataset, k-fold cross validation is used to evaluate the forecasts, with 31 forecasts per fold.

This method of assessing the relationship between forecast skill and dataset length is not perfect due to the temporally varying precipitation climatology. For example, one potential issue is that the smallest datasets, which have fewer folds, get verified only against testing data from the same season as the training data. As more data are added, the size of the training set increases, but the training set starts to include data from other times of the year relative to the test set. Therefore, it is possible that these "new" training data add only limited value to each testing fold. Additionally, the uncertainty of the forecast itself changes with time due to seasonal variations in climatology, such that, as more dates are added to the dataset, the overall

forecast difficulty (and thus, objective skill) changes depending on what dates are added. Despite these deficiencies, the results give useful preliminary insight into the feasibility of adopting ML-based techniques operationally.

3. Results

a. Traditional verification metrics over the full domain

1) ROC METRICS

All forecasts have good discrimination ability, as indicated by ROC diagrams (Figs. 5a,d,g,j,m,p,s,v) and AUC (Figs. 6a–d). Even the worst-performing forecast system (i.e., the raw SREF ensemble for the 3-in. threshold; Fig. 4d) has an AUC of 0.80. Nevertheless, for all thresholds (all but the 3-in. threshold), the SREF (HREFv2) RFFPs have significantly greater AUC than the corresponding raw and smoothed ensemble probabilities ($p < 0.0001$; Figs. 7a,c,e,g). The SREF RFFPs also have significantly greater AUC than the raw HREFv2 probabilities ($p < 0.0001$; Figs. 7a,c,e,g).

Interestingly, the raw SREF forecast probabilities often have greater AUC compared to the raw HREFv2 forecast probabilities, even though the HREFv2 is a CAE that performs subjectively better than the SREF. This behavior likely reflects the insensitivity of the AUC to bias (thus negating the SREF's poor reliability; e.g., Figs. 5b,e,h,k,n,q and 6i–k) as well as the larger membership of the SREF, which enables the raw SREF to issue more unique forecast probabilities and thus have more unique ''points'' on its ROC curve, possibly increasing AUC.

2) RELIABILITY

The raw SREF and HREFv2 probabilities suffer from substantial overforecasting bias at all precipitation thresholds, with the raw SREF forecasts generally having the worst reliability (Figs. 5b,e,h,k,n,q,t,w and 6i–l). The 0.1-in. raw SREF forecasts (Fig. 5b) have particularly poor reliability, as the reliability curve falls below the no skill line for multiple forecast probability bins. Meanwhile, the raw HREFv2 reliability curves contain ''gaps'' (Figs. 5e,k,q,w) since, with only 8 members, the HREFv2 is unable to issue probabilities in all bins. Spatially smoothing the raw ensemble forecasts improves reliability and removes the gaps from the raw HREFv2 reliability curves (Figs. 5b,e,h,k,n,q,t,w). The RF technique tends to produce even better (i.e., near perfect) forecast reliability for both ensembles at most thresholds (Figs. 6i–l).

3) PERFORMANCE DIAGRAMS

Performance diagrams suggest that the skill of the RFFPs matches or exceeds that of the other sets of forecasts at all four precipitation thresholds (Figs. 5c,f,i,l,o,r,u,x). The SREF RFFPs clearly outperform corresponding raw and smoothed SREF forecasts (Figs. 5c,i,o,u), while the HREFv2 RFFPs have the greatest relative performance at the 0.1-in. threshold (Fig. 5f). At the other thresholds (Figs. 5l,r,x), the HREFv2 RFFPs and smoothed probabilities demonstrate similar skill, which noticeably exceeds that of the raw HREFv2 probabilities.

One interesting characteristic of the SREF performance diagrams (Figs. 5c,i,o,u) is that the second-best performing probabilities (in terms of CSI) tend to be from the raw SREF (e.g., Figs. 5c,i,o). This is because the smoothed SREF probabilities require a relatively large amount of spatial smoothing to optimize the BS (Fig. 3a), and this degrades resolution (Figs. 6m–p). Hence, for the SREF forecasts, a main advantage of the RF technique is that it calibrates the raw ensemble probabilities while improving—rather than sacrificing—resolution.

4) BSS AND BS COMPONENTS

With only one exception (i.e., the smoothed 3-in. HREFv2 probabilities), the RFFPs have significantly greater BSSs ($p < 0.0001$) than the corresponding raw and smoothed ensemble probabilities (Figs. 7b,d,f,h). At the 0.1-in. threshold, the SREF RFFPs even have a significantly greater BSS than the raw HREFv2 probabilities ($p < 0.0001$; Fig. 7b), which is remarkable given the much coarser horizontal grid spacing of the SREF. The RF-based approach improves the BSS by simultaneously enhancing forecast reliability and resolution (Figs. 6e,i,m).

The RFFPs provide the greatest increase in BSS relative to the corresponding raw and smoothed ensemble forecasts at the smallest precipitation thresholds (Figs. 6e–h), likely because the smallest thresholds have the greatest climatological frequency (Fig. 8). More occurrences of yes observations in the training dataset make it easier for the RF to identify the systematic relationships between the predictors and observations.

RFFPs always have better resolution than the corresponding raw and smoothed ensemble forecast probabilities (Figs. 6m–p) and nearly always have better reliability (Figs. 6i–l). It is also noteworthy that the RFFPs increase resolution relative to the spatially smoothed ensemble forecasts, both in cases where the two-dimensional spatial smoothing technique degrades

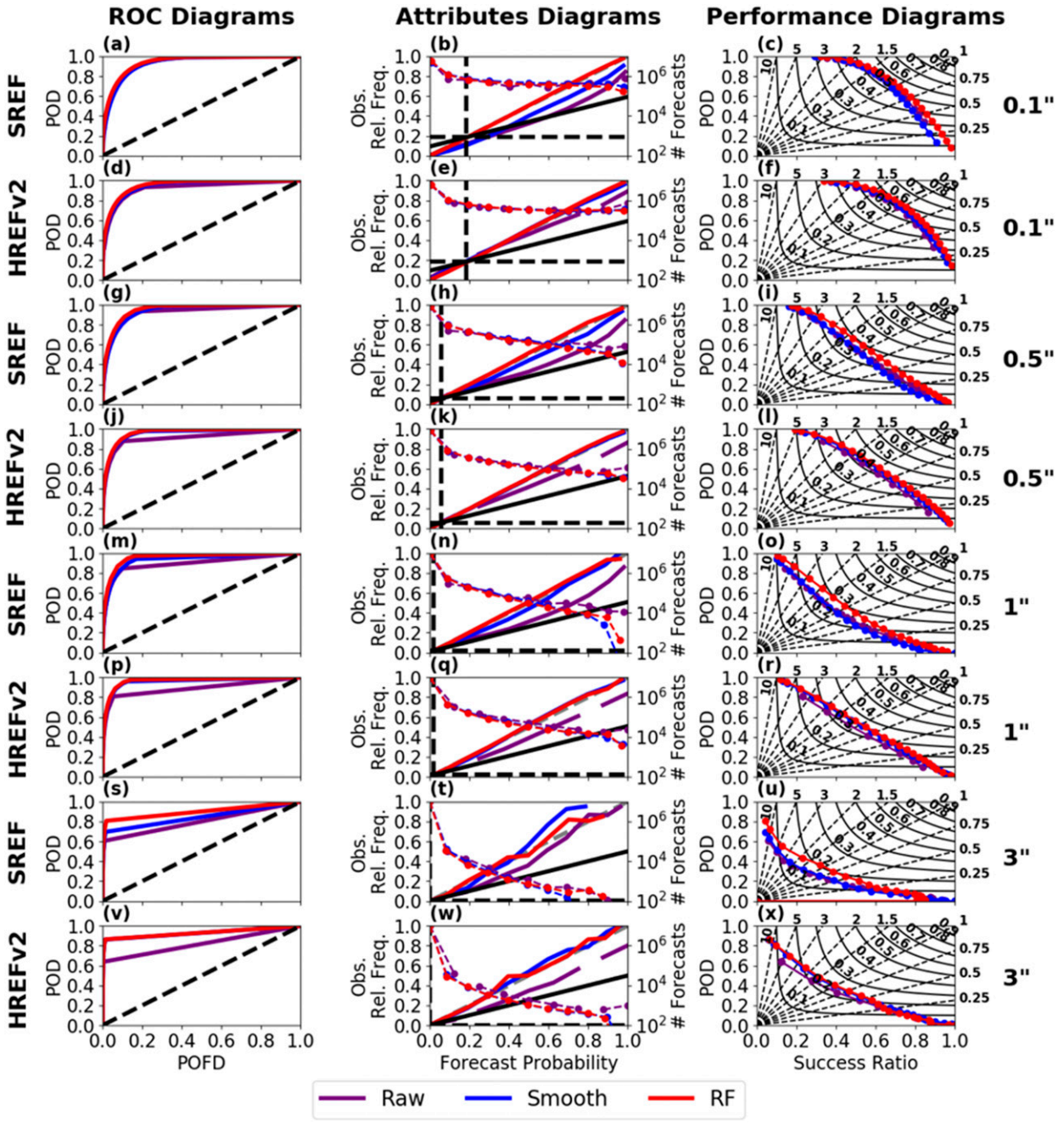

FIG. 5. (a) ROC curve for the SREF at the 0.1-in. threshold for raw (purple), smooth (blue), and RF (red) forecasts. The black dashed line indicates a random forecast. (b) Attributes diagram for the SREF at the 0.1-in. threshold for the same forecasts as in (a). Black dashed lines indicate the relative frequency of the sample climatology, the solid black line is the ''no skill'' line, and the dashed gray line represents perfect reliability. The number of forecasts in each probability bin is indicated by the colored dashed lines with filled circles. (c) Performance diagrams for the SREF at the 0.1-in. threshold for the same forecasts as in (a). Lines of constant bias are dashed, while lines of constant CSI are solid. Each of 21 forecast probability levels are indicated by filled circles. (d)–(f) As in (a)–(c), but for the HREFv2. (g)–(l) As in (a)–(f), but for the 0.5-in. threshold. (m)–(r) As in (a)–(f), but for the 1-in. threshold. (s)–(x) As in (a)–(f), but for the 3-in. threshold.

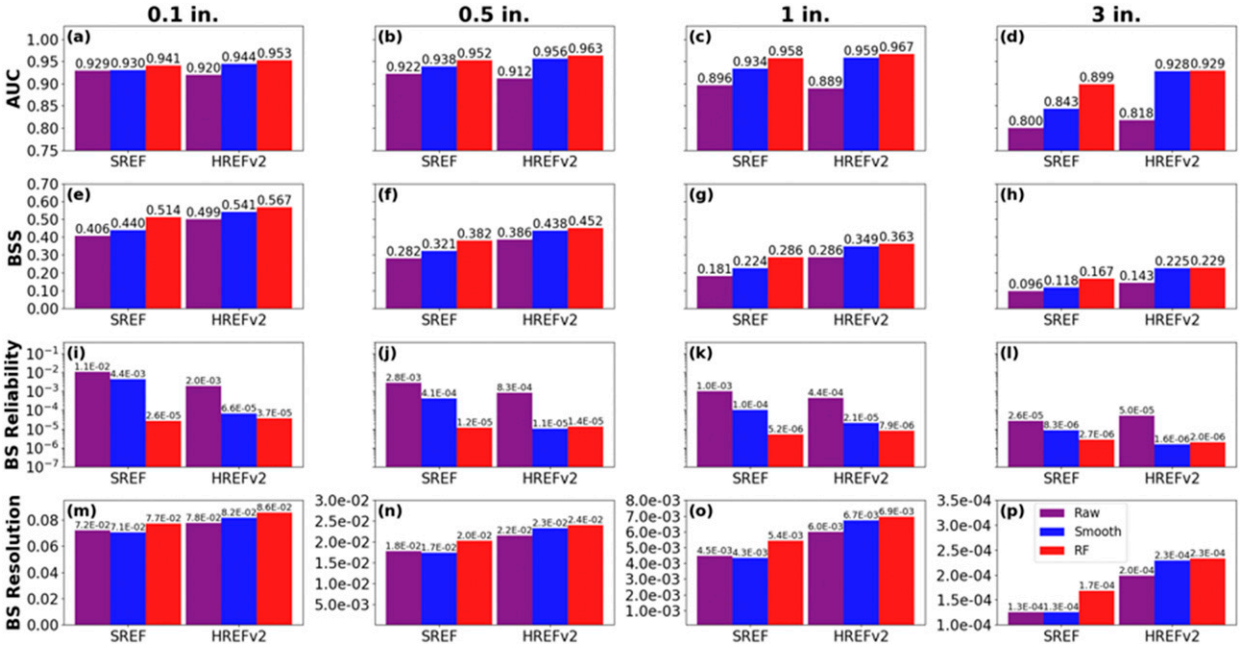

FIG. 6. (a) AUC for SREF and HREFv2 raw (purple), smooth (blue), and RF forecasts (red) for the 0.1-in. threshold. (b)–(d) As in (a), but for the 0.5-, 1-, and 3-in. thresholds, respectively. (e)–(h) As in (a)–(d), but for BSS. (i)–(l) As in (a)–(d), but for the reliability component of the BS. (m)–(p) As in (a)–(d), but for the resolution component of the BS. Note the different *y* axes for (m)–(p), and note that lower values of BS reliability are better.

(e.g., the SREF forecasts) and enhances (e.g., the HREFv2 forecasts) reliability.

b. *Regional results*

Similar results are obtained when forecasts are verified regionally. For the SREF, the RF-based approach improves the BSS in every region at every threshold compared to the raw and smoothed ensemble forecasts (Figs. 9a–d). These greater BSSs can be attributed to both better reliability and resolution (Figs. 9a–d). Importantly, the RF approach appears to improve the BSS and BS components approximately equally for each region at each threshold (with a few exceptions; e.g., the West region benefits disproportionately at the 1-in. threshold). This finding suggests that a single, CONUS-wide RF can learn enough spatial information such that the benefits to RF-based postprocessing are not confined to a single region.

The same general findings also apply to the HREFv2: at each threshold, each region benefits from the RF-based postprocessing approximately equally (Figs. 10a–d). Of course, these benefits are most pronounced for the lower thresholds, consistent with the full-domain findings presented above. Regardless, the results suggest that, for a given threshold, a single, CONUS-wide RF can provide reliability and resolution benefits to forecasts in all regions, despite each region having different climatological frequencies of threshold exceedance (e.g., Figs. 9 and 10).

c. *Full-domain spatial biases*

Full-domain spatial bias magnitudes are small for both ensembles, as the center of the observed conditional distribution seldom falls more than 20–40 km from the yes forecast point (Figs. 11a–x). The spatial biases are greatest for the raw and smoothed SREF forecasts (Figs. 11a,b,g,h,m,n,s,t) and for the higher (i.e., 1 and 3 in.; Figs. 11m–x) precipitation thresholds. These findings make sense given that the higher thresholds are more likely to be associated with deep convection, which is more difficult to predict—especially for a convection-parameterizing ensemble (e.g., Kain et al. 2006)—due to uncertainties in initiation and evolution. The anisotropy of the conditional distribution of observed yes events seen in Figs. 11a–x is consistent with Marsh et al. (2012), who obtained a similar preferred southwest–northeast orientation and explained that it reflects the mean shape and orientation of individual precipitation objects over the full dataset.

One important finding in the present study is that the RF technique helps alleviate spatial biases in the raw and smoothed ensemble probabilities. This result can be seen in two distinct ways. First, the center of the distribution (i.e., the red dot in Figs. 11a–x) is closest

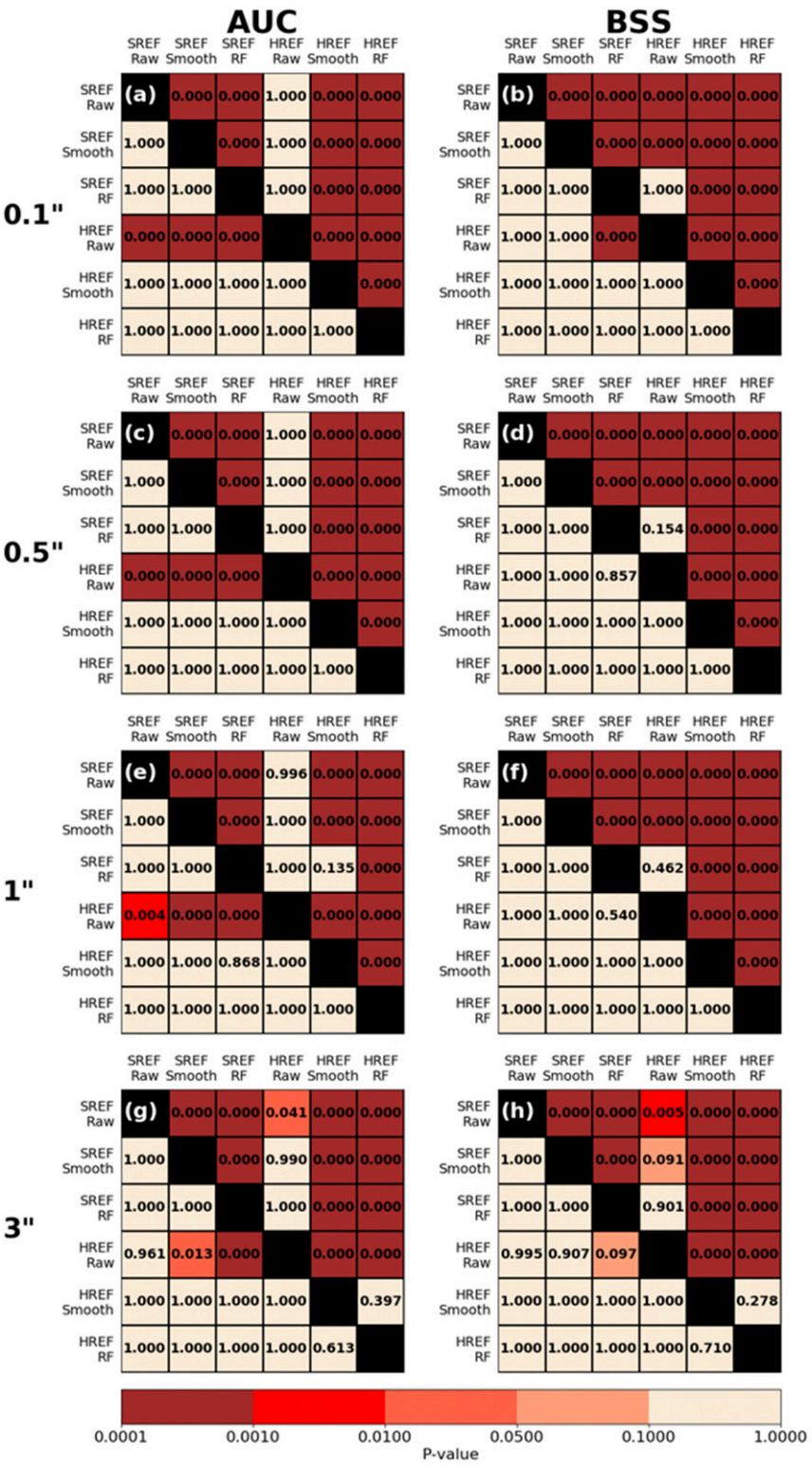

FIG. 7. (a) The *p* values from one-sided paired permutation significance tests for AUC for the 0.1-in. threshold. (b) As in (a), but for BSS. (c),(d) As in (a) and (b), but for the 0.5-in. threshold. (e),(f) As in (a) and (b), but for the 1-in. threshold. (g),(h) As in (a) and (b), but for the 3-in. threshold. Each square reports the *p* value associated with testing whether the forecast displayed across the top row has a significantly greater metric than that from the forecast displayed along the left-hand column.

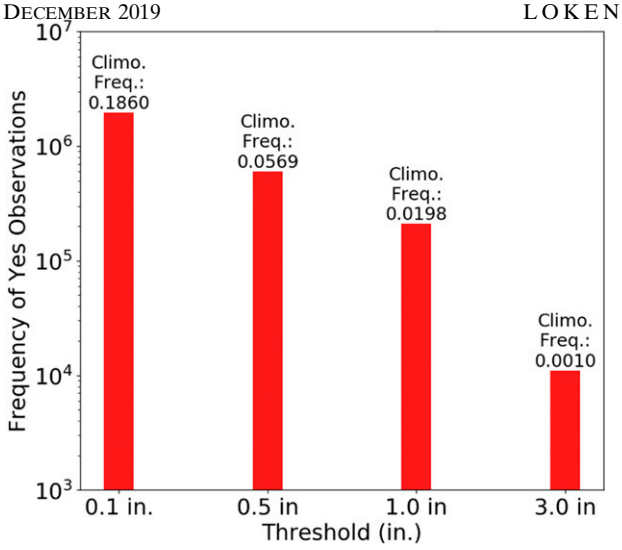

FIG. 8. Number of "yes" observations (i.e., instances when the observed 24-h precipitation exceeds the given threshold) at the 0.1-, 0.5-, 1-, and 3-in. thresholds. The corresponding relative frequency, abbreviated as "Climo. Freq.," is displayed above each bar. Note the logarithmic *y* axis.

to the yes forecast point (i.e., the black dot in Figs. 11a–x) in the RF plots (i.e., Figs. 11c,f,i,l,o,r,u,x). Additionally, difference plots (Figs. 12a–p) show that the RF technique tends to add conditional observations in locations that oppose the direction of the spatial bias and/or subtract conditional observations from locations in the same direction of the spatial bias. For example, in the 1-in. raw and smoothed SREF forecasts, the center of the observed distribution falls too far to the southeast of the yes forecast point (Figs. 11m,n). In both cases, the RF technique adds conditional observations to the northwest and subtracts conditional observations to the southeast (Figs. 12i,j) so that the center of the RF-based conditional distribution of observed yes events is closer to the yes forecast point (Fig. 11o). Similar behavior is seen for both ensembles at all thresholds, although the effect is stronger for the SREF since the HREFv2 forecasts have fewer spatial biases. In many cases, the RF approach also adds conditional yes observations to the yes forecast point and surrounding points (e.g., Figs. 12g,n,o), which improves the forecast by increasing the conditional probability of a yes observation given a yes forecast.

d. Sensitivity of results to dataset length

The best AUC and BSS values are generally obtained using a dataset of 248 days (Figs. 13a,b,d,e). Interestingly, increasing the dataset beyond 248 days results in slightly lower AUCs and BSSs. This finding can potentially be explained by temporal variations in the observed precipitation climatology. For example, since AUC is sensitive to the number of correct negatives, AUC may be artificially inflated (deflated) during times of the year with lower (higher) forecast uncertainty. Indeed, this is exactly the pattern that is seen (Figs. 13a,c,d,f). The temporal variation in climatology may also help explain the behavior of the 3-in. SREF and HREFv2 BSS curves, which reach a local minimum at 372 days. Although difficult to discern from Figs. 13c and 13f, the 3-in. uncertainty reaches a minimum (maximum) at 372 (124) days. A relatively low (high) forecast uncertainty makes a reference forecast of climatology more (less) skillful and more (less) harshly penalizes small forecast errors. Thus, the BSS decreasing after 124 (248) days for the SREF (HREFv2) may be at least partly explained by the variations in the already-low observed precipitation climatology.

Because these variations in climatology have the potential to "artificially" influence the verification metrics, the results should be interpreted cautiously. Nevertheless, it is likely that the results presented herein are not due entirely to temporal variations in the dataset climatology, especially since the BSS follows a similar pattern as AUC. For both AUC and BSS, there are obvious gains from increasing the length of the dataset from 62 to 124 days and, in general, additional gains from further increasing the dataset to 248 days. Since each fold's testing set contains 31 days, these findings suggest that a minimum training set length of 93–217 days (i.e., approximately 1–2 seasons) is desirable for adequate performance.

e. Select cases

Two cases are subjectively selected to illustrate the RFFPs' relative performance on individual days.

1) 1200 UTC 2 OCTOBER–1200 UTC 3 OCTOBER 2017

The heaviest precipitation during this period occurred in a corridor extending from northeastern Minnesota into west-central Kansas ahead of a cold front. Relatively heavy precipitation also occurred in northern Montana downstream of a midlevel shortwave trough, while southern Louisiana and southern Florida experienced weakly forced tropical showers.

The raw SREF and HREFv2 probabilities performed relatively well at all four thresholds (Figs. 14a,d,g,j,m,p,s,v). In general, these probabilities had good sharpness and resolution. However, these raw ensemble forecasts also placed 90%–100% probabilities in locations where the observed precipitation did not exceed the threshold (e.g., southern Utah in Figs. 14a,d). The spatially smoothed forecasts (Figs. 14b,e,h,k,n,q,t,w) helped calibrate the raw forecast probabilities but had reduced

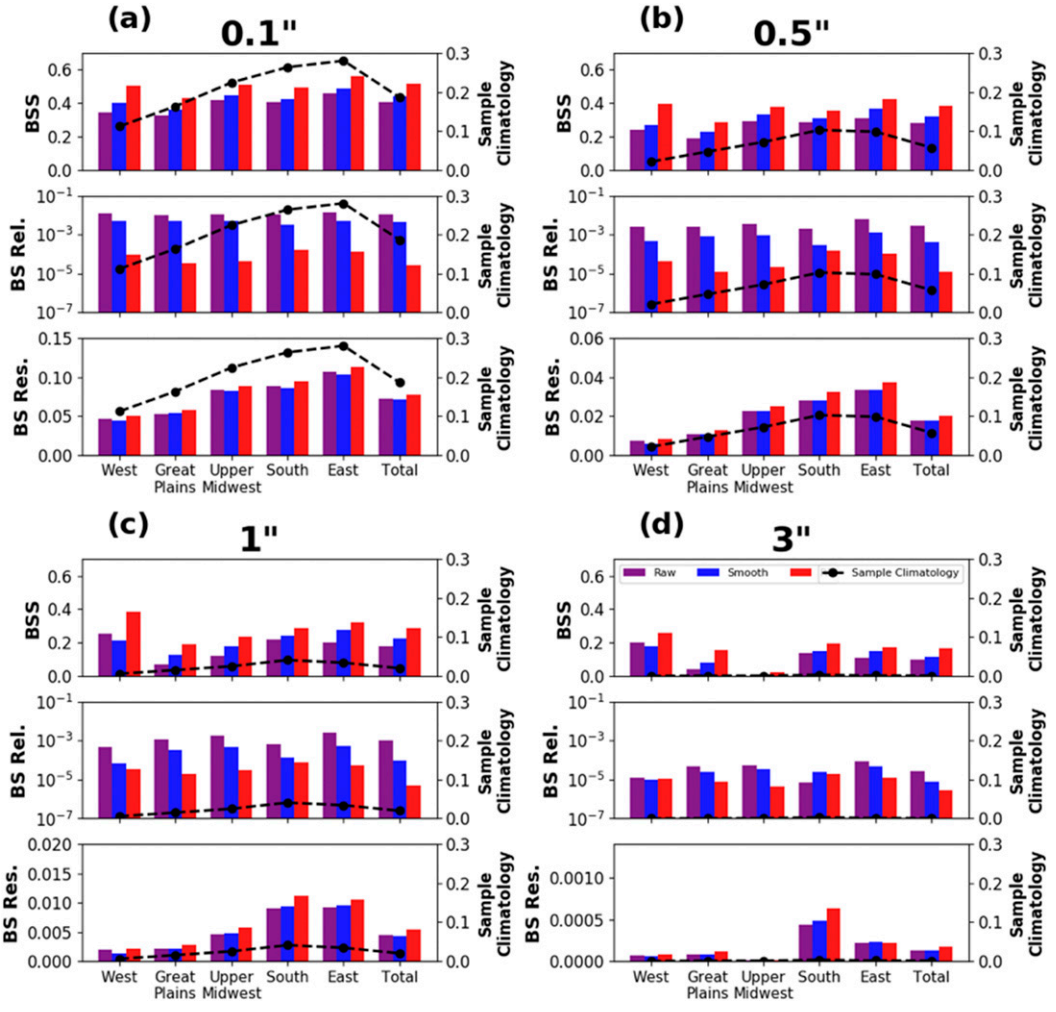

FIG. 9. Regional BSS, BS reliability, and BS resolution for the raw (purple), spatially smoothed (blue), and RF-based (red) SREF forecasts at the (a) 0.1-, (b) 0.5-, (c) 1-, and (d) 3-in. thresholds. In each case, the black dashed line indicates the climatological relative frequency of threshold exceedance in the given region. Full domain metrics are also given under the "Total" label.

sharpness. Meanwhile, the RFFPs (Figs. 14c,f,i,l,o,r,u,x) generally had good calibration, sharpness, and resolution. For example, like the 0.5-in. raw SREF probabilities (Fig. 14g), the 0.5-in. SREF RFFPs (Fig. 14i) exceeded 80% over east-central Minnesota and northern Montana, while the spatially smoothed SREF probabilities (Fig. 14h) were less in both areas. Moreover, the 0.5- and 1-in. SREF RFFPs (Figs. 14i,o) had less false alarm area over the High Plains compared to the spatially smoothed SREF forecasts (Figs. 14h,n). Differences between the HREFv2 smoothed probabilities and corresponding RFFPs were subtler since less spatial smoothing was required to calibrate the raw HREFv2 probabilities. For example, compared to the corresponding smoothed forecasts (Figs. 14k,q), the 0.5- and 1-in. HREFv2 RFFPs (Figs. 14l,r) had a larger spatial extent of >90% probabilities in the Upper Midwest where observed precipitation exceeded the threshold. The 0.5-in. RFFPs (Fig. 14l) also gave slightly lower probabilities in east-central South Dakota but slightly enhanced the probabilities in central Iowa compared to the spatially smoothed probabilities (Fig. 14k).

2) 1200 UTC 22 JUNE–1200 UTC 23 JUNE 2017

Early in this period, elevated storms were ongoing over South Dakota, Minnesota, Wisconsin, and Michigan. Later, surface-based storms formed ahead of a cold front extending from eastern Ontario into central Kansas and eastern Colorado, bringing heavy rainfall to southern Wisconsin, central Michigan, and northern New York. Eastern Colorado and western Kansas also experienced

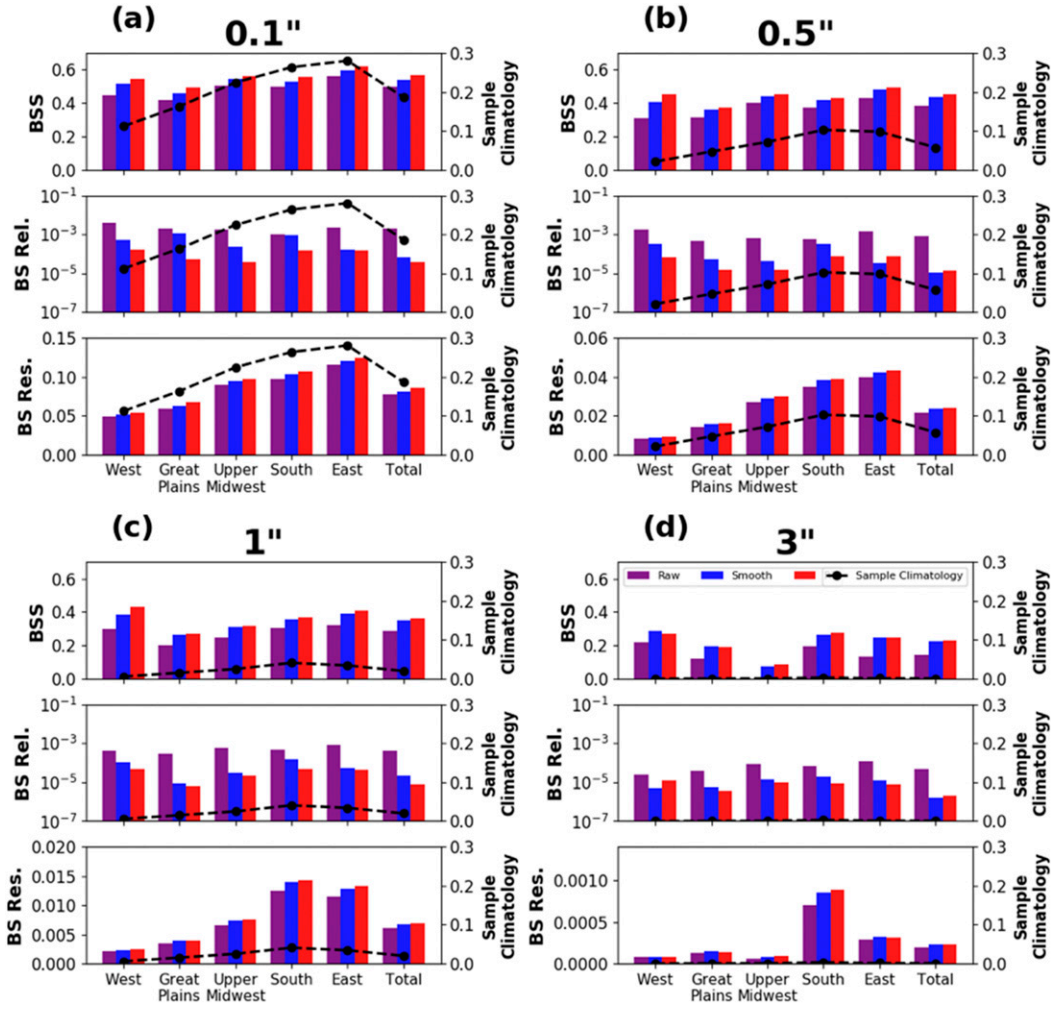

FIG. 10. As in Fig. 9, but for the HREFv2 forecasts. Axes are the same as in Fig. 9.

0.1–0.5-in rainfall associated with postfrontal upslope flow. Meanwhile, Tropical Storm Cindy brought heavy rainfall to the southeastern United States.

Raw ensemble probabilities from the SREF and HREFv2 (Figs. 15a,d,g,j,m,p,s,v) predicted the day's precipitation relatively well, despite several instances of overconfidence (e.g., central Colorado, northeastern Mississippi, and eastern California in Fig. 15a; extreme southwestern Kentucky in Fig. 15j) and misses (e.g., northwestern Nebraska in Fig. 15d; southern Iowa in Fig. 15m). Spatially smoothing the raw ensemble probabilities (Figs. 15b,e,h,k,n,q,t,w) generally helped improve calibration and POD, but forecasts remained imperfect. For example, 0.1-in. SREF exceedance probabilities (Fig. 15b) remained near 1 in southwestern Kentucky and northeastern Mississippi, while the 0.1-in. HREFv2 smoothed probabilities over northwestern Nebraska remained less than 2%. The RFFPs (Figs. 15c,f,i,l,o,r,u,x) tended to fix these problems. The 0.1-in. SREF-based RFFPs gave smaller probabilities in northeastern Mississippi (Fig. 15c), while the 0.1-in. HREFv2-based RFFPs gave higher (i.e., 2%–10%) probabilities in northwestern Nebraska. In general, the RFFPs (Figs. 15c,f,i,l,o,r,u,x) had good calibration, sharpness, and resolution. They tended to increase POD and sharpness compared to the spatially smoothed forecasts while only modestly increasing POFD. For example, the HREFv2 1-in. RFFPs (Fig. 15r) gave higher probabilities in northern Alabama compared to the raw (Fig. 15p) and smoothed (Fig. 15q) HREFv2 forecasts while the false alarm area increased only slightly. Similarly, the SREF-based 3-in. RFFPs had better POD in central Alabama (Fig. 15u) with a false alarm area only slightly greater than the corresponding raw and smoothed ensemble forecasts (Figs. 15s–t). While the RFFPs did not always improve on the raw and smoothed ensemble probabilities (e.g., central Michigan in Figs. 15v–x), the general performance of the RFFPs was strong.

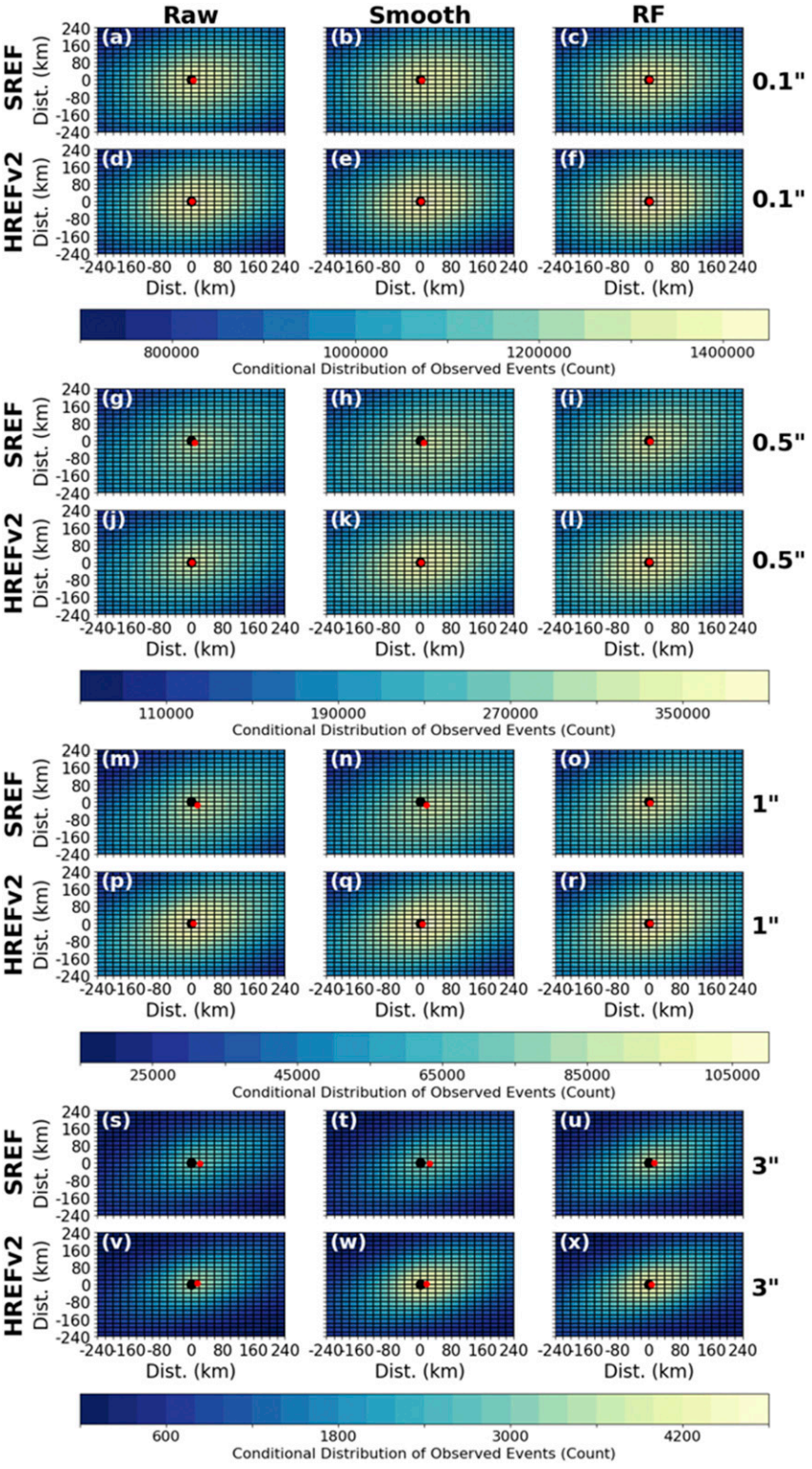

FIG. 11. (a) Spatial distribution of observed yes events given a yes forecast (see text) at point (0, 0) (black dot) for the raw SREF ensemble forecast at the 0.1-in. threshold. The red dot denotes the center of the distribution. (b),(c) As in (a), but for the SREF-derived smoothed and RF-based forecasts. (d)–(f) As in (a)–(c), but for the HREFv2. (g)–(l) As in (a)–(f), but for the 0.5-in. threshold. (m)–(r) As in (a)–(f), but for the 1-in. threshold. (s)–(x) As in (a)–(f), but for the 3-in. threshold. Note the different color scale used for each threshold.

4. Summary and discussion

This paper describes a technique to postprocess ensemble probabilistic precipitation forecasts year-round over the contiguous United States (CONUS) using a single random forest (RF). Specifically, the RF-based postprocessing is applied to 24-h (1200 UTC–1200 UTC) probabilistic precipitation forecasts from the Short-Range Ensemble Forecast System (SREF; Du et al. 2015) and the High-Resolution Ensemble Forecast System, Version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019) at four precipitation thresholds: 0.1 in. (2.54 mm), 0.5 in. (12.7 mm), 1 in. (25.4 mm), and 3 in. (76.2 mm). Random forest forecast probabilities (RFFPs) are compared against each ensemble's raw probabilities (i.e., the fraction of members exceeding a threshold) and spatially smoothed probabilities (i.e., raw ensemble probabilities smoothed in space using an isotropic two-dimensional Gaussian kernel density function to optimize the Brier score).

Relative to these baseline forecasts, the RFFPs provide better reliability and resolution, fewer spatial biases, and statistically greater Brier skill scores (BSSs) and areas under the relative operating characteristics curve (AUCs). The RFFPs perform best at lower thresholds, which have greater climatological frequencies and thus provide more examples of "yes observations" for the algorithm to use to discern data patterns associated with threshold exceedance. The RF-based postprocessing also benefits the SREF more than the HREFv2, a result that makes sense given that the raw SREF contains more systematic biases than the raw HREFv2. The result may also indicate that different ensembles require different sets of predictor variables to achieve the best postprocessing benefits. For example, it is possible that, for the HREFv2, the ensemble mean is not as meaningful as an ensemble summary characteristic as it is for the SREF. Similarly, it is possible that the HREFv2 forecast variables contain more small-scale noise than those from the SREF because of the HREFv2's finer horizontal grid spacing.

The biggest advantage of the RFFPs is that they provide a convenient "summary" product that is calibrated with respect to forecast probability magnitudes and spatial coverage. While near-perfect reliability can also be achieved using two-dimensional spatial smoothing with the proper value of σ, spatially smoothing ensemble probabilities reduces sharpness (e.g., Sobash et al. 2011, 2016; Loken et al. 2017, 2019) and potentially sacrifices resolution if too much smoothing is required. Moreover, the "best" value of σ may vary based on geographic location and time of year (e.g., Fig. 3), as precipitation uncertainty is reduced where stronger and/or more predictable forcing is present, such as near high terrain (e.g., Blake et al. 2018) or during the cold season (e.g., Schwartz et al. 2019). Thus, while a time- and space-varying σ may be required to properly calibrate forecasts using spatial smoothing, the RF-based approach implicitly accounts for spatial and temporal variations in precipitation uncertainty.

In practice, RFFPs could provide value to forecasters as an ensemble summary product that would eliminate the need for internal forecaster calibration of ensemble biases. Indeed, the RFFPs would fill an important operational need by quickly conveying reliable uncertainty information to the forecaster (Evans et al. 2014). The RFFPs could also be used as an automated "first guess" probabilistic precipitation forecast field, which could increase forecaster efficiency (e.g., Karstens et al. 2018). Importantly, the implementation of RFFPs into operations would be computationally feasible. While training RFs can be expensive, particularly when many predictor variables and training examples are used, using a trained RF to make real-time predictions is cheap. For example, real-time RFFPs are currently being generated from 0000 UTC HREFv2 data. Including the preprocessing step, the RFFPs can be made in 30 min or less on a single processor.

Nevertheless, ML-based postprocessing has several important drawbacks. Most notably, since ML-based techniques "learn" based on past results, they require quality historical datasets of sufficient length for both the forecast and observations. When modifications are made to the ensemble forecast system, it is often advisable to retrain the RF with forecast data from the new system, since, while the underlying statistical relationships between the forecast and observed variables may generally hold, the optimal splitting thresholds in the RF may change as biases enter or exit the ensemble system. It is an open question (and probably situation dependent) whether the RF can be retrained simply by adding the new forecast data to the training set (along with the old data) or if the RF should be retrained entirely "from scratch" using only the new data. Fortunately, even if the RF requires retraining from scratch, preliminary results herein suggest that a training set of "only" 93–217 days is required to create skillful RFFPs; nevertheless, even 93 days represents a substantial gap between the implementation of the new system and the ability to create skillful RFFPs. Moreover, due to the reduced observed climatological frequency of the higher threshold exceedances, it may be necessary to have more data for the RFFPs to outperform spatially smoothed ensemble probabilities at the highest thresholds (e.g., 3 in. and greater), which tend to be most impactful in terms of their threat to life and property.

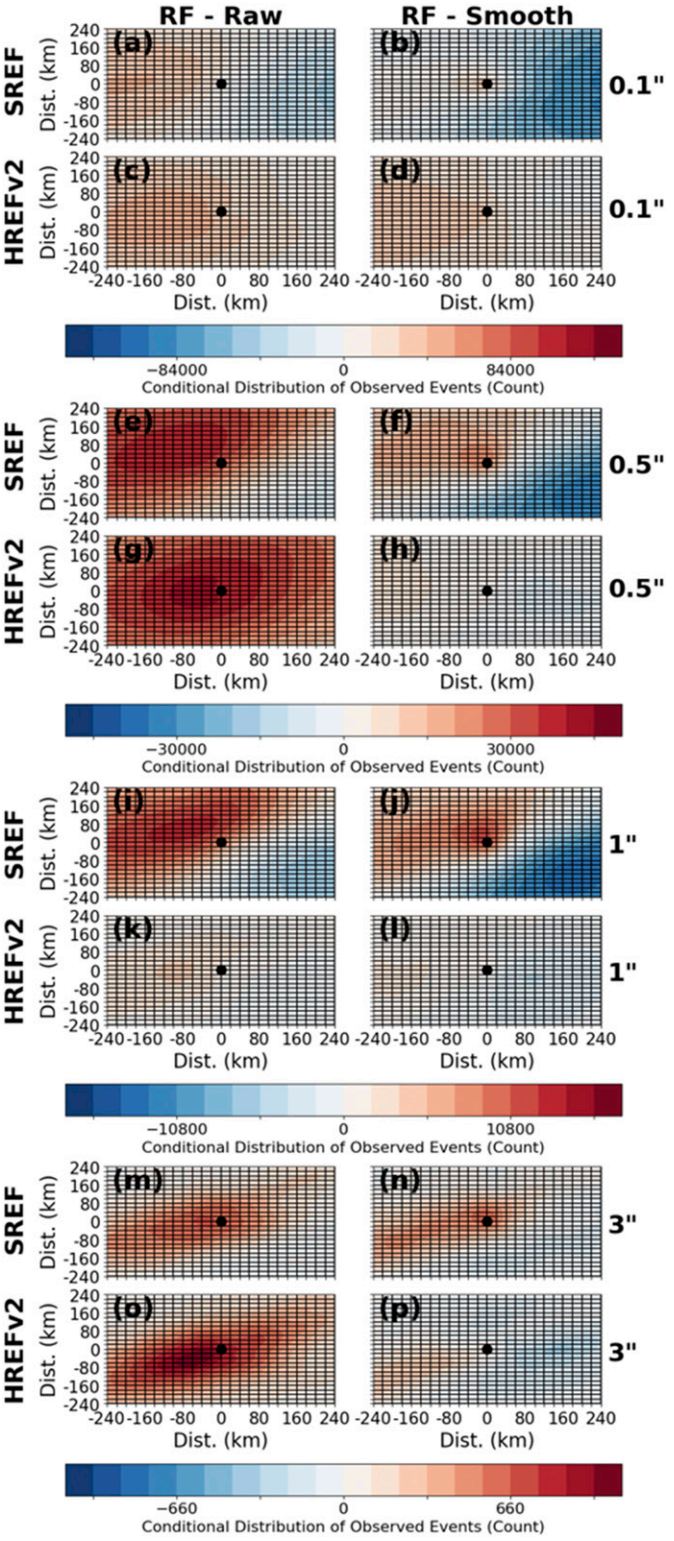

FIG. 12. (a) Difference between the conditional distribution of yes observed events given a yes SREF-based RF forecast at point (0, 0) (black dot) and the conditional distribution of yes observed events given a yes raw SREF forecast at (0, 0) at the 0.1-in. threshold (i.e., Fig. 11c minus Fig. 11a). (b) As in (a), but subtracting the smoothed SREF

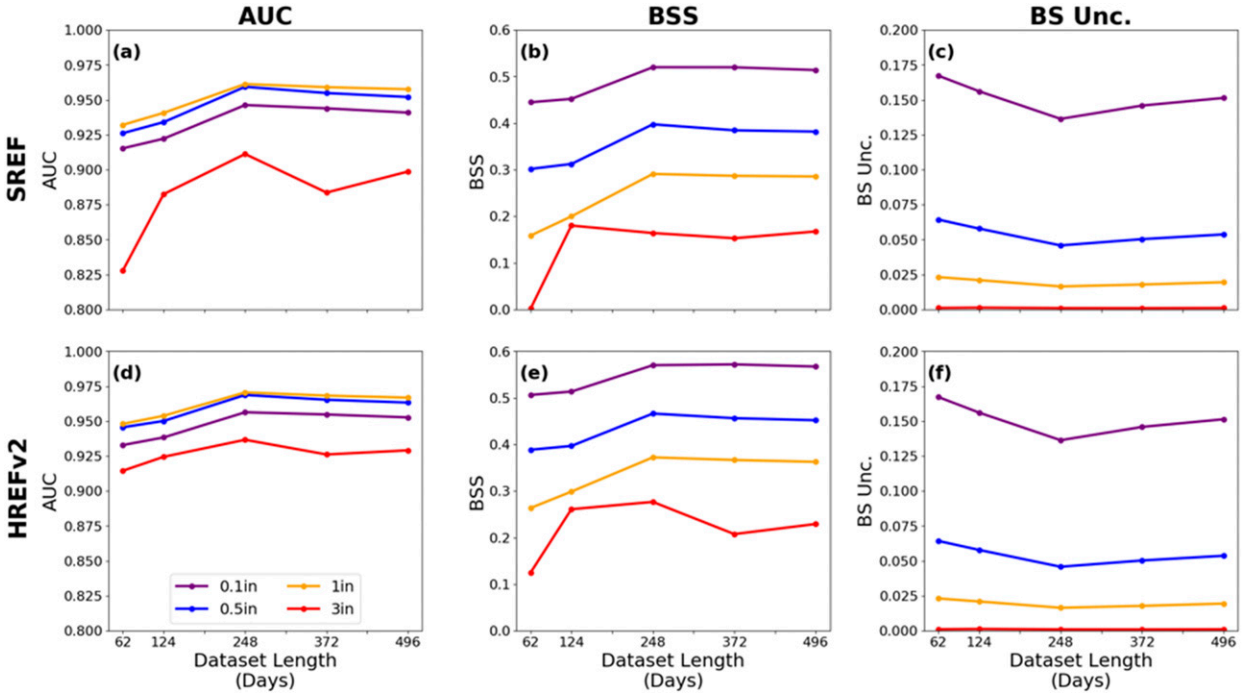

FIG. 13. (a) AUC as a function of dataset length for the SREF. (b),(c) As in (a), but for the BSS and uncertainty component of the BS, respectively. (d)–(f) As in (a)–(c), but for the HREFv2.

Another drawback of the RF-based approach is that the RFFPs are not *always* superior to raw or spatially smoothed ensemble probabilities at every location during every day, and it can be difficult to determine where and why the ML algorithm struggles, particularly in the absence of interpretability information [e.g., partial dependence plots and individual conditional expectation plots (Goldstein et al. 2015); variable importance (McGovern et al. 2017)]. Therefore, developing and applying useful ML interpretability metrics is an important topic of ongoing research (e.g., Gagne et al. 2019; Herman and Schumacher 2018a). Another important limitation of ML compared to other postprocessing techniques is that it can require a substantial degree of hyperparameter tuning to produce a skillful forecast. Moreover, there are no formal guidelines for constructing the ML model itself, and it can be impossible to know if the model being used is designed optimally. Finally, as with other postprocessing techniques, the skill of the RFFPs will ultimately be related to and limited by the skill of the underlying dynamical model (e.g., Gagne et al. 2014). Therefore, while ML-based postprocessing techniques can serve as useful tools, they do not eliminate the need for human forecasters and model developers.

5. Conclusions and future work

As computing storage and resources continue to increase, opportunities to effectively apply ML to meteorological datasets will undoubtedly become more numerous as well. This paper provides a first attempt at addressing some basic considerations regarding the utilization of machine learning for NWP postprocessing. Despite the drawbacks associated with ML-based postprocessing, it is found that RFFPs can provide calibrated probabilistic precipitation forecasts whose quality matches or exceeds that of spatially smoothed ensemble probabilities. Indeed, it is promising that a single RF can attain such forecast quality, especially

←

distribution from the SREF RF distribution (i.e., Fig. 11c minus Fig. 11b). (c),(d) As in (a) and (b), but for the HREFv2. (e)–(h) As in (a)–(d), but for the 0.5-in. threshold. (i)–(l) As in (a)–(d), but for the 1-in. threshold. (m)–(p) As in (a)–(d), but for the 3-in. threshold.

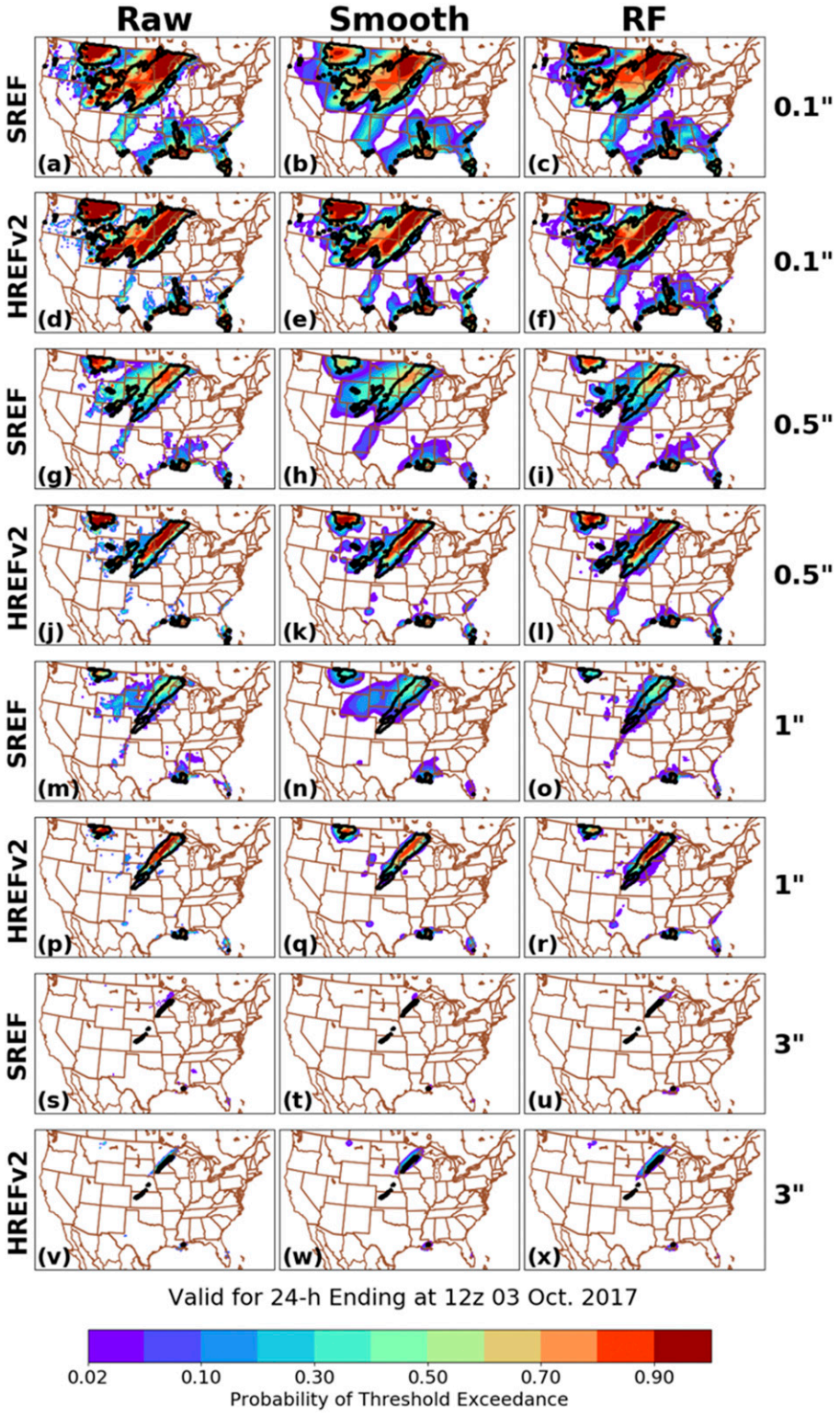

FIG. 14. (a) The 0.1-in. PQPFs from the raw SREF ensemble, valid for the 24-h ending at 1200 UTC 3 Oct 2017. The black contours indicate where the observed precipitation exceeded the given threshold. (b),(c) As in (a), but for the spatially smoothed and RF-based SREF PQPFs. (d)–(f) As in (a)–(c), but for the HREFv2. (g)–(l) As in (a)–(f), but for the 0.5-in. threshold. (m)–(r) As in (a)–(f), but for the 1-in. threshold. (s)–(x) As in (a)–(f), but for the 3-in. threshold.

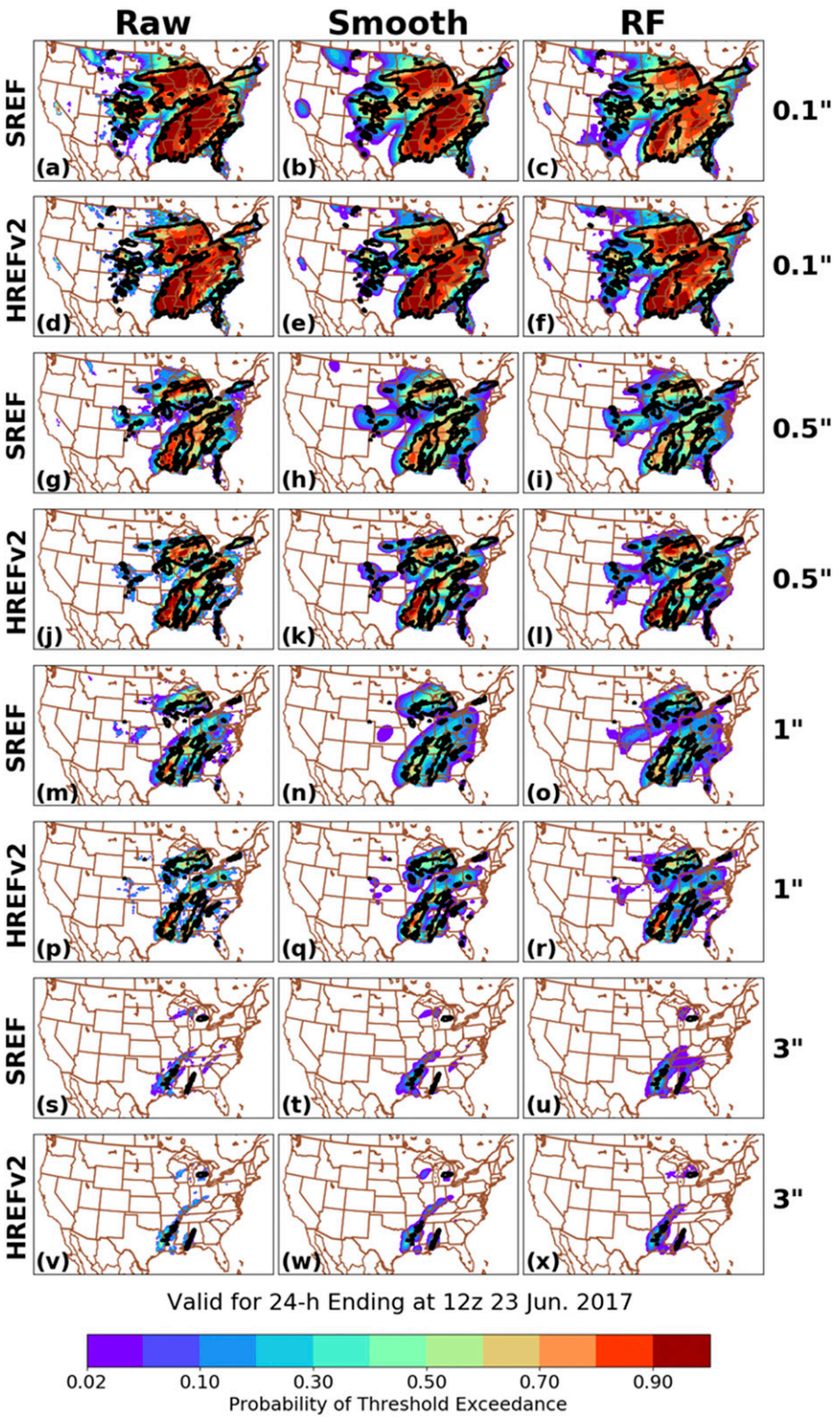

FIG. 15. As in Fig. 14, but for the 24-h period ending at 1200 UTC 23 Jun 2017.

given the relatively simplistic RF design and short (i.e., <1.5 years) dataset.

Future work should explore using more complex ML-based techniques for postprocessing and/or other RF constructions. For example, in the present study, individual-member forecasts were initially used as predictors, but this implementation consumed too much memory to be feasible. However, if variable importance and/or feature selection (e.g., McGovern et al. 2017; Herman and Schumacher 2018c) were used to strategically reduce the number of predictor variables, predictors from more sources could potentially be incorporated into the algorithm. Including interpretability metrics (e.g., partial dependence plots or individual conditional expectation plots; Goldstein et al. 2015) may also provide value to forecasters using the product in real time. Given that the precipitation climatology over the CONUS varies in space and time (Schumacher and Johnson 2006), using separate RFs for individual regions and seasons may add further interpretability and skill to the RFFPs. Other ML methods, such as deep learning, may produce better RFFPs and enhance interpretability as well. Because this study examined the impacts of ML-based postprocessing on ad hoc, multimodel, multiphysics ensembles, future work should investigate how ML-based postprocessing affects other, more formally designed ensembles (e.g., the NCAR Ensemble; Schwartz et al. 2015, 2019). Finally, future work may wish to apply the general methods of this study to other prediction problems, such as severe weather, forecasting for longer or shorter time periods, and summarizing ensemble output from multiple NWP sources. It is also recommended that current and future products be evaluated in an operational setting, such as the Flash Flood and Intense Rainfall Experiment (Albright and Perfater 2018) or the NOAA Hazardous Weather Testbed Spring Forecasting Experiment (e.g., Gallo et al. 2017) to more directly assess value to forecasters.

Acknowledgments. Support for this work was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Additional support was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. Stage IV precipitation data were provided by NCAR/EOL under the sponsorship of the National Science Foundation. The Stage IV data were accessed from https://data.eol.ucar.edu/. We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Some of the computing for this project was also performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU).

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Adrianto, I., T. Trafalis, and V. Lakshmanan, 2009: Support vector machines for spatiotemporal tornado prediction. *Int. J. Gen. Syst.*, **38**, 759–776, https://doi.org/10.1080/03081070601068629.

Albright, B., and S. Perfater, 2018: 2018 flash flood and intense rainfall experiment. Weather Prediction Center Rep., 96 pp., https://www.wpc.ncep.noaa.gov/hmt/2018_FFaIR_final_report.pdf.

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, https://doi.org/10.1175/MWR-D-17-0277.1.

Baldwin, M. E., J. S. Kain, and S. Lakshmivarahan, 2005: Development of an automated classification procedure for rainfall systems. *Mon. Wea. Rev.*, **133**, 844–862, https://doi.org/10.1175/MWR2892.1.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.*, **112**, 677–691, https://doi.org/10.1002/qj.49711247307.

Blake, B. T., J. R. Carley, T. I. Alcott, I. Jankov, M. E. Pyle, S. E. Perfater, and B. Albright, 2018: An adaptive approach for the calculation of ensemble gridpoint probabilities. *Wea. Forecasting*, **33**, 1063–1080, https://doi.org/10.1175/WAF-D-18-0035.1.

Breiman, L., 1984: *Classification and Regression Trees*. Wadsworth International Group, 358 pp.

——, 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the Southwest monsoon. *Wea. Forecasting*, **17**, 1080–1100, https://doi.org/10.1175/1520-0434(2002)017<1080:SREFOP>2.0.CO;2.

Bukovsky, M. S., 2011: Masks for the Bukovsky regionalization of North America. Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, accessed 1 August 2019, http://www.narccap.ucar.edu/contrib/bukovsky/.

Carley, J. R., B. R. J. Schwedler, M. E. Baldwin, R. J. Trapp, J. Kwiatkowski, J. Logsdon, and S. J. Weiss, 2011: A proposed model-based methodology for feature-specific prediction for high-impact weather. *Wea. Forecasting*, **26**, 243–249, https://doi.org/10.1175/WAF-D-10-05008.1.

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

——, W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, ——, and M. L. Weisman, 2010a: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, ——, M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, https://doi.org/10.1175/2010MWR3624.1.

——, and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427, https://doi.org/10.1175/2009WAF2222258.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

——, G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. Preprints, *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5, https://ams.confex.com/ams/27WAF23NWP/webprogram/Manuscript/Paper273421/NWP2015_NCEP_RegionalEnsembles_paper.pdf.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

Epstein, E. S., 1969: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8**, 190–198, https://doi.org/10.1175/1520-0450(1969)008<0190:TROIUI>2.0.CO;2.

Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, https://doi.org/10.1175/WAF-D-13-00064.1.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta Model. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 10.1, http://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm.

Gagne, D. J., II, A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

——, A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2019: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Wea. Forecasting*, **34**, 151–164, https://doi.org/10.1175/WAF-D-18-0108.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin, 2015: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.*, **24**, 44–65, https://doi.org/10.1080/10618600.2014.907095.

Good, P. I., 2006: *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser, 228 pp.

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.*, **121**, 764–787, https://doi.org/10.1175/1520-0493(1993)121<0764:PEOAUB>2.0.CO;2.

Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345, https://doi.org/10.1175/1520-0434(1999)014<0338:PFUANN>2.0.CO;2.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, https://doi.org/10.1175/WAF-D-10-05038.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

——, and ——, 2018a: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

——, and ——, 2018b: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeor.*, **19**, 1753–1776, https://doi.org/10.1175/JHM-D-18-0092.1.

——, and ——, 2018c: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141**, 4564–4575, https://doi.org/10.1175/MWR-D-12-00297.1.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

——, 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

——, and R. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB). Part 1 Dynamics. NCAR Tech. Note NCAR/TN-489+STR, 75 pp., https://doi.org/10.5065/D6WH2MZX.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., P9.137, https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.

——, C. J. Melick, and S. J. Weiss, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, https://ams.confex.com/ams/28SLS/webprogram/Session41668.html.

——, A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

——, S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, https://doi.org/10.1175/WAF906.1.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Karstens, C. D., and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Key, J., J. Maslanik, and A. Schweiger, 1989: Classification of merged AVHRR and SMMR Arctic data with neural networks. *Photogramm. Eng. Remote Sensing*, **55**, 1331–1338.

Kishtawal, C. M., S. Basu, F. Patadia, and P. K. Thapliyal, 2003: Forecasting summer rainfall over India using genetic algorithm. *Geophys. Res. Lett.*, **30**, 2203, https://doi.org/10.1029/2003GL018504.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682, https://doi.org/10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2.

Kuligowski, R. J., and A. P. Barros, 1998: Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Wea. Forecasting*, **13**, 1194–1204, https://doi.org/10.1175/1520-0434(1998)013<1194:LPFFAN>2.0.CO;2.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Lin, Y., 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data, version 1.0. UCAR/NCAR Earth Observing Laboratory, accessed 15 April 2019, https://data.eol.ucar.edu/dataset/21.093.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

——, ——, ——, and ——, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

Malone, T., 1955: Application of statistical methods in weather prediction. *Proc. Natl. Acad. Sci. USA*, **41**, 806–815, https://doi.org/10.1073/pnas.41.11.806.

Manzato, A., 2007: Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecasts. *Atmos. Res.*, **83**, 349–365, https://doi.org/10.1016/j.atmosres.2005.10.021.

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. M. Hitchens, and J. Hardy, 2012: A method for calibrating

deterministic forecasts of rare events. *Wea. Forecasting*, **27**, 531–538, https://doi.org/10.1175/WAF-D-11-00074.1.

Marzban, C., and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar–derived attributes. *J. Appl. Meteor.*, **35**, 617–626, https://doi.org/10.1175/1520-0450(1996)035<0617:ANNFTP>2.0.CO;2.

Mason, S. J., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada Level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

NCEI, 2019: U.S. billion-dollar weather and climate disasters: Overview. NOAA/NCEI, https://www.ncdc.noaa.gov/billions/.

Nielsen, E. R., and R. S. Schumacher, 2018: Dynamical insights into extreme short-term precipitation associated with supercells and mesovortices. *J. Atmos. Sci.*, **75**, 2983–3009, https://doi.org/10.1175/JAS-D-17-0385.1.

Ortiz-García, E. G., S. Salcedo-Sanz, and C. Casanova-Mateo, 2014: Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmos. Res.*, **139**, 128–136, https://doi.org/10.1016/j.atmosres.2014.01.012.

Palmer, T., 2017: The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *J. Adv. Model. Earth Syst.*, **9**, 730–734, https://doi.org/10.1002/2017MS000999.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Rajendra, P., K. V. N. Murthy, A. Subbarao, and R. Boadh, 2019: Use of ANN models in the prediction of meteorological data. *Model. Earth Syst. Environ.*, **4**, 1051–1058, https://doi.org/10.1007/s40808-019-00590-2.

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. Kain, 2019: Post-processing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

——, D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Schumacher, R. S., and R. H. Johnson, 2006: Characteristics of U.S. extreme rain events during 1999–2003. *Wea. Forecasting*, **21**, 69–85, https://doi.org/10.1175/WAF900.1.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, Z. Liu, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

——, G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

——, ——, K. Fossell, R. Sobash, and M. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. *Mon. Wea. Rev.*, **145**, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's Real-Time Convection-Allowing Ensemble Project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127**, 433–446, https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2.

Szpiro, G. G., 1997: Forecasting chaotic time series with genetic algorithms. *Phys. Rev. E*, **55**, 2557–2568, https://doi.org/10.1103/physreve.55.2557.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747, https://doi.org/10.1175/1520-0493(2001)129<0729:EOASRM>2.0.CO;2.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.*, **8**, 209–219, https://doi.org/10.1017/S1350482701002092.

Wong, K. Y., C. L. Yip, and P. W. Li, 2008: Automatic tropical cyclone eye fix using genetic algorithm. *Expert Syst. Appl.*, **34**, 643–656, https://doi.org/10.1016/j.eswa.2006.10.013.

Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B. http://ams.confex.com/ams/pdfpapers/124587.pdf.